THE UNIVERSITY OF CHICAGO

THE GENETIC ARCHITECTURE OF COMPLEX TRAITS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON CANCER BIOLOGY

BY
KESTON AQUINO-MICHAELS

CHICAGO, ILLINOIS
DECEMBER 2015

to my wife and family for their encouragement and love

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The concerted effort to characterize the contribution of inherited variation to human health and disease through genome wide association studies (GWAS) has promised to increase our understanding of the diverse molecular pathways underlying specific human traits, as well as yield clinically actionable findings that would prove useful in estimating disease risk and designing personalized therapeutic approaches. However, it has become evident in recent years that further work is needed in order to harness the biological and translational potential of GWAS findings.

One approach to glean further insight from GWAS stems from the seminal discovery that trait-associated SNPs are enriched for expression quantitative trait loci (eQTL), highlighting an important mechanism by which SNPs influence traits, i.e. through modulation of gene transcript levels. Similarly, valuable insight into biological mechanisms can be gained through expanding eQTL studies to include genetic regulation of non-coding RNA levels, whose inter-individual variation, like gene expression, has been shown to impact on complex diseases.

Here we identified a large collection of distant miRQTL (391) and replicated with 26% $p < 0.05$ and 98.5% allelic concordance (67 of 68 SNPs) in an independent cohort. Analysis of genomic properties of replicated miRQTL reveal strong enrichment for mapping within ENCODE-annotated functional elements. *In-silico* analysis using replicated distant miRQTL reveal local mRNA expression quantitative trait loci (eQTL) putatively regulating microRNA abundance. Mediation analysis and association testing between microRNA and mRNA confirms *HEXIM1* as a putative regulator of hsa-mir-185-5p levels. These results highlight a potential novel mechanism of long-range regulation of microRNA abundance, providing valuable insight into the biology underlying complex traits.

A second approach to transform GWAS findings into clinically relevant tools is based on developing and applying statistical prediction methods to existing GWAS information. These

approaches have the potential to translate genetic data into clinically relevant predictions of risk. It is these high-confidence predictions of complex traits such as disease risk or drug response that will fulfill the goal of personalized medicine.

Therefore we proposed a novel systems approach to complex trait prediction, which leverages and integrates similarity in genetic, transcriptomic or other omics-level data. Using seven disease datasets from the Wellcome Trust Case Control Consortium (WTCCC), we show that OmicKriging has important translational potential. In addition, we built a statistical-learning machine, which integrates large-scale whole-genome data to predict bevacizumab-induced hypertension in cancer patients. We found that incorporating primary genetic as well as clinical trial data into our model significantly improves prediction and therefore should motivate the use of such large-scale whole-genome predictors in a clinical setting. Taken together, these approaches utilize novel methodology as well as publicly available datasets to yield valuable mechanistic insight into genetic regulation of complex traits, and provide genetic tools for clinical implementation.

CHAPTER 1

INTRODUCTION

The principal goal of genetic medicine is to provide biological insight which will lead to improved patient care by uncovering genetic mechanisms that underlie human phenotypic variation. Many phenotypes follow simple patterns of genetic inheritance, as discovered by Gregor Mendel and are more recently studied using family based models[58]. Notable successful family based studies have identified genes linked to human diseases, ranging from cystic fibrosis to breast cancer[6, 36]. However many common traits including various cancer types, type 2 diabetes and height cannot not be explained solely by simple patterns of inheritance[43]. Known as complex traits, they typically involve environmental factors and complex genetic architectures that make the process of elucidating their genetic mechanisms particularly challenging.

Since the completion of the human genome project, a comprehensive catalog of genetic polymorphisms has enabled researchers to systematically identify genome-wide complex trait-associated-loci through genome wide association studies (GWAS)[113, 55]. GWAS have enjoyed a tremendous success having identified 19,603 trait-associated-loci to date[118]. Furthermore GWAS have revealed that often complex traits are influenced by many genetic loci of small effect rather than a single locus of large effect as is characteristic of Mendelian traits[72]. GWAS have also highlighted key mechanistic differences between Mendelian and complex trait-associated-loci by showing that unlike protein-altering Mendelian loci, the majority of complex trait-associated-loci are intronic or intergenic[118, 23]. This motivated the field of human genetics to investigate the genetic mechanisms underlying complex traits, which resulted in the foundational understanding that complex trait-associated-loci are more likely to be gene expression quantitative trait loci (eQTL)[78]. This finding suggested that complex trait-associated-loci influence traits by altering gene expression levels, which ultimately alter phenotypes.

In contrast to simple Mendelian traits, the identification of complex trait-associated-loci yielding clinically actionable information has been far more limited. An exception is the field of pharmacogenomics, which has been successful at identifying trait-associated-loci of modest effect[121, 14]. However replication of pharmacogenomic trait-associated loci remains limited[50] These associated loci often influence drug response through biological pathways involved in drug action or metabolism and therefore yield mechanistic insight as well as clinical impact[121]. By comparison prevention and early treatment of common complex diseases such as cancer and type 2 diabetes aided by GWAS have been far less effective. This is largely due to trait-associated-loci accounting for a small proportion of the phenotypic variance. For example, type-2 diabetes-associated loci account for only $\sim$10% of the phenotypic variance. Conversely twin and family studies suggest that the proportion of phenotypic variance accounted for by additive genetic effects should be significantly larger. This discrepancy is know as the problem of missing heritability[72]. Larger GWAS sample sizes may improve this estimate, for example Park et al suggest that very large samples ($\sim$ 1 million) will be required to close this gap[84]. However with recent development of powerful statistical learning tools[17, 1], clinically relevant patient predictions despite small GWAS sample sizes are becoming a reality.

The work presented in this thesis is connected through the goal of advancing our understanding of clinically relevant biology through genetic and statistical methodologies. To that end, this thesis is comprised of individual chapters that specifically focus on the identification of genetic mechanisms that may ultimately benefit patient care, the development of novel methodologies for genetic medicine, and the applications of existing statistical learning methodologies motivating their use in the clinic. Chapter 2 focuses on microRNA quantitative trait loci mapping in the Framingham Heart Study cohort with replication in Geuvadis cohort. Chapter 3 develops the extension of Kriging for genetic medicine. Chapter 4 applies statistical learning methods to predict bevacizumab induced hypertension in clinical trials

cohorts.

Our current understanding of the genetic mechanisms underlying regulation of microRNA in humans is limited. To date only a handful of studies have identified miRQTL loci, the vast majority of which have been local associations. While previous studies have reported miRQTL, none have been replicated and little focus has been on distant genetic regulation. To address these gaps in knowledge we conducted miRQTL mapping using the Framingham Heart Study cohort (n=5134) individuals using 180 microRNA and 2 million HapMap 2 imputed genotypes. We identified a large collection of distant miRQTL (391) and replicated with 26% $p < 0.05$ and 98.5% allelic concordance (67 of 68 SNPs) in an independent cohort. Analysis of genomic properties of replicated miRQTL revealed strong enrichment for mapping within ENCODE-annotated functional elements compared to an empirical null distribution. *In-silico* analysis using replicated distant miRQTL revealed local mRNA expression quantitative trait loci (eQTL) putatively regulating microRNA abundance. Mediation analysis and association testing between microRNA and mRNA confirmed *HEXIM1* as a putative regulator of hsa-mir-185-5p levels. These results highlight a potential novel mechanism of long-range regulation of microRNA abundance, providing valuable insight into the biology underlying complex traits, discussed in Chapter 2.

Chapter 3 focuses on the systematic integration of high-throughput biological technologies for clinical application by predicting individual patient disease risk. Specifically, in chapter 3 we extend a statistical framework, Kriging to the setting of genetic medicine. Kriging is a method from geospatial statistics that interpolates missing measurements by distance weighted average of the surrounding observations. We show that this approach has strong clinical potential when combined with several sources of high-throughput data (microRNA, mRNA, genotype). Furthermore we show that the method can be universally and flexibly applied to complex traits of varying genetic architectures.

The motivation behind this approach was to develop a approach to predicting complex

traits using multiple levels of omic data while simultaneously remaining fast and scalable. This approach was contrasted to several Bayesian approaches to complex trait prediction which attempted to learn polygenic and sparse genomic architectures. The Bayesian approaches often provided excellent prediction performance when restricted to SNP genotype data and immune mediated traits. Nevertheless these approaches typically used algorithms that required tremendous compute times to converge.

This prompted us to develop a flexible, scalable and universal tool for prediction of complex traits. Intuitively our method OmicKriging interpolates a missing data point, for example rainfall at a given weather station by weighted average of the rainfall at the surrounding weather stations. In this example, the weight is determined by euclidean distance to the missing weather station. In OmicKriging, we translate this concept to the goal of predicting a missing phenotype by weighted average of the other phenotypes in the cohort. Analogously, our average is weighted by distance similarity by genotype. Importantly, we show that we can achieve comparable results to Bayesian approaches with runtime on the order of minutes rather than hours. Our approach using OmicKriging is also external-data-integrative as we show that substantial prediction performance can be achieved by allowing genome-wide significant loci identified in previous studies to take equal weight with the rest of the genome. Finally, we provide an R package that allows researches to easily apply OmicKriging.

In Chapter 4 we focus on developing methodologies to predict bevacizumab induced hypertension also referred to as provocative hypertension. Bevacizumab in a monoclonal antibody that inhibits VEGF formation. It is used for the treatment of several cancer types including colorectal, lung, renal, ovarian, and glioblastoma multiforme. Grade 2/3 hypertension is a common side effect seen in approximately 15% of patients undergoing bevacizumab therapy. Interestingly Saif et al found that pre-existing hypertension does not predispose patients to grade 2/3 hypertension seen with bevacizumab treatment[93]. Therefore in this

chapter we focus on applying existing methodologies that have had tremendous success in the statistical/machine learning community to improve our understanding and demonstrate clinically relevant prediction of bevacizumab induced hypertension using clinical trial cohorts.

Here we specifically utilized clinical trials cohorts CALGB 80303 and 90401. CALGB 80303 is a randomized double-blind, placebo-controlled phase III trial of gemcitabine with bevacizumab versus gemcitabine with placebo, while CALGB 90401 is a randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer. Importantly we show that the use of external sources of data provide relevant feature information for prediction.

To do this we used XC-Pleiotropy primary hypertension data to predict grade 2/3+ provocative hypertension in CALGB 80303 (AUC = 0.61 for grade 3+). This performance result is comparable to prediction of primary hypertension observed in Chapter 3. We built a statistical learning machine using CALGB 90401 as training data to predict in CALGB 80303. We show that the LASSO, a statistical method ideally suited for high-dimensional underdetermined systems of equations, is capable of significant prediction performance in CALGB 80303 (AUC = 0.6 for grade 3+). We show that an ensemble approach which combines LASSO with an additional statistical learning approach called random forest improves prediction in CALGB 80303 (AUC=0.68). Finally we show that the combination of models built with CALGB 90401 and XC-Pleiotropy primary hypertension data perform significantly better than either alone (AUC=0.71).

Therefore we highlight that large scale meta-analysis of primary hypertension cohorts are directly and non-redundantly capable of significantly predicting hypertension and improving prediction performance of hypertension in combination with other sources of data. We show clinical relevant prediction performance (AUC = 0.71) which should motivate the use of

statistical learning approaches to identify and monitor patients at risk of hypertension while undergoing bevacizumab therapy. Importantly we highlight the secondary utility of the vast amount of GWAS level data in improving clinically relevant predictions.

CHAPTER 2

THE GENETIC LANDSCAPE OF DISTANT MICRORNA REGULATION

Abstract

Distant genetic regulation of microRNA remains poorly understood, with only a handful of distant microRNA quantitative trait loci (miRQTL) reported, none of which have been replicated in an independent cohort. Here we identify a large collection of distant miRQTL (391) and replicate 26% at $p < 0.05$, with 98.5% allelic concordance (67 of 68 SNPs) in an independent cohort. Analysis of genomic properties of replicated miRQTL reveal strong enrichment for mapping within ENCODE-annotated functional elements. *In-silico* analysis using replicated distant miRQTL reveal local mRNA expression quantitative trait loci (eQTL) putatively regulating microRNA abundance. Mediation analysis and association testing between microRNA and mRNA confirms the RNA polymerase II transcription inhibitor *HEXIM1* as a putative regulator of hsa-mir-185-5p levels. These results highlight a potential novel mechanism of long-range regulation of microRNA abundance, providing valuable insight into the biology underlying complex traits.

Introduction

The concerted effort to characterize the contribution of inherited variation to human health and disease through genome wide association studies (GWAS) has promised to increase our understanding of the diverse molecular pathways underlying specific human traits. However, it has become evident in recent years that further work is needed in order to harness the biological and translational potential of GWAS findings.

One approach to glean further insight from GWAS stems from the seminal discovery that

trait-associated SNPs are enriched for expression quantitative trait loci (eQTL) [78, 119, 126, 4], highlighting an important mechanism by which SNPs influence traits, i.e. through modulation of gene transcript levels. Valuable insight into biological mechanisms can be gained through expanding eQTL studies to investigate genetic regulation of non-coding RNA levels, whose inter-individual variation, like gene expression, has been shown to impact on human health and disease[38, 66, 64, 97].

MicroRNAs are small non-coding RNAs that play important gene-regulatory roles by sequence-specific pairing to the mRNAs of protein-coding genes to direct their post transcriptional repression [2, 3, 57]. As such, microRNAs coordinate normal developmental and physiological processes [8, 52, 9], including cellular proliferation, differentiation and apoptosis, and their aberrant expression or alteration contributes to a range of human pathologies, including cancer [68, 82, 39].

Several microRNA QTL-mapping studies have identified genetic variants associated with microRNA expression [5, 85, 30, 10, 59, 46]. Many of these studies reported that microRNA QTL may constitute eQTLs for target gene transcripts [30, 59], and identified several microRNA QTL as trait-associated variants [30, 46], suggesting that microRNA QTL mapping is a valuable tool to dissect genetic mechanisms underlying gene regulation. However, the identification of distant microRNA QTL has only begun to be investigated, with few associations reported [59, 30, 5], likely owing to small sample sizes and the multiple-testing burden. Furthermore, evidence for replication of distant microRNA remains elusive. The identification and replication of distant eQTLs as important regulators of gene expression in other studies motivates a similar endeavor for microRNA QTL [119].

In this study, we identify the largest collection of microRNA QTL to date, with over a quarter of distant associations replicating with near perfect concordance. We further characterize the genomic features of microRNA QTL, as well as microRNA-mRNA interactions and reveal a potential novel mechanism of long-range microRNA regulation.

Results

Genome-wide miRQTL analysis

To improve statistical power to detect microRNA-associated variants (miRQTL), we performed a miRQTL mapping analysis in 5,135 peripheral blood samples from the Framingham Heart Study (FHS), and replication analysis on lymphoblastoid cell lines derived from 331 individuals of European descent in the Geuvadis Consortium. Several methods that estimate global confounding factors in eQTL studies have been convincingly shown to increase the number of reproducible eQTL associations[62, 99]. These methods build on factor analysis approaches to identify and subsequently correct for variation other than genotype. The more recently developed probabilistic estimation of expression residuals (PEER) factor analysis identifies a small number of factors for each sample and assumes each factor is relevant across the transcriptome[100]. Therefore to correct for confounding sources of variation, we incorporated PEER factor analysis.

In the discovery dataset we identified 1,665 SNP-microRNA associations (distant and local) corresponding to 1,228 SNPs and 79 microRNA (FDR < 0.05). Of these 1,665 we identified 391 distant (trans-chromosome) miRQTL (FDR < 0.05) (Figure 2.1A-B and Supplementary Table 1), corresponding to 356 unique SNPs and 53 unique microRNAs. We observed well-calibrated type I error rates with a genomic control factor of 1.02 (Figure 2.1A). We verified that top association results remain substantively unchanged when mixed effects models were used to adjust for family-relatedness.

To assess reproducibility of the detected distant miRQTL, we replicated distant miRQTL from our discovery cohort in our independent Geuvadis dataset (n=331). Despite tissue differences and vastly smaller sample size, over a quarter (26.5%, $p < 0.05$) of distant associations replicated (Figure 2.2C-D and Supplementary Table 2), with 98.5% (67 of 68) showing concordant allelic direction of effect in the replication cohort compared with the discovery

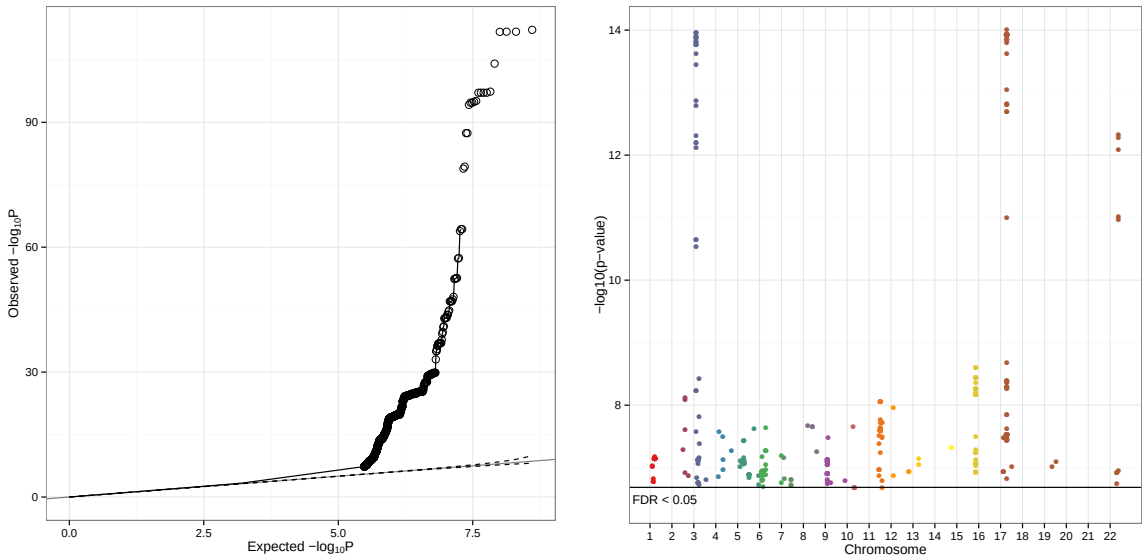

Figure 2.1: **Discovery of distant miRQTL** A) QQ-plot of all tested SNP-MircroRNA pairs in FHS. We observed well controlled genomic inflation ($\lambda = 1.02$) B) Manhattan plot of FDR < 0.05 distant miRQTL in FHS.

analysis (Supplementary Figure 1). These results suggest that distant genetic regulation of microRNA expression is far more widespread than previously recognized (Borel 2011, Huan 2015).

We examined the genomic properties of the identified distant-miRQTL variants utilizing ENCODE annotation. We found that these SNPs were significantly enriched for mapping within protein binding sites (ChIP-seq) and open chromatin (DNAse-seq, FAIRE) (p = 5e-3, p = 3e-3 respectively) (Fig. 3A-B). Additionally we found using HaploReg that miRQTL were enriched for being located within myeloid derived K562 cells specific histone enhancers regions (fold change 5; Fig. 3C), compared to the enhancer regions observed in three non-blood cell lines[117]. Enhancer enrichment within myeloid cells is consistent with detection of blood-derived distant miRQTL.

Given that these SNPs were enriched for mapping within functional elements, we hypothesized that distant miRQTL influence expression of microRNA through modulation of expression levels of a local transcript, whose gene product may influence microRNA abun-

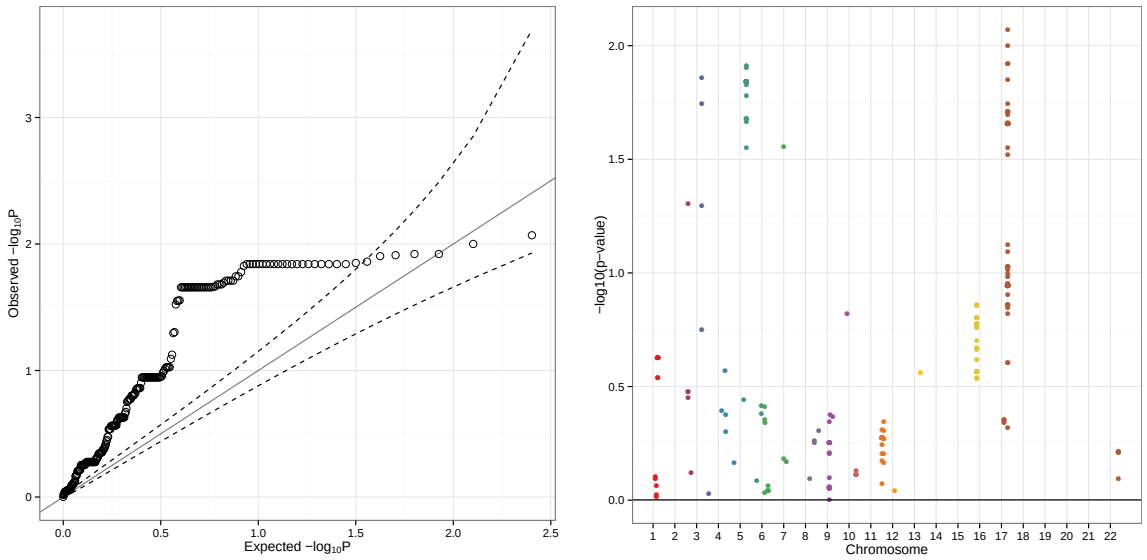

Figure 2.2: **Replication of distant miRQTL results** A) QQ-plot of all tested SNP-MircroRNA pairs identified in the discovery cohort. B) Manhattan plot of p < 0.05 distant miRQTL.

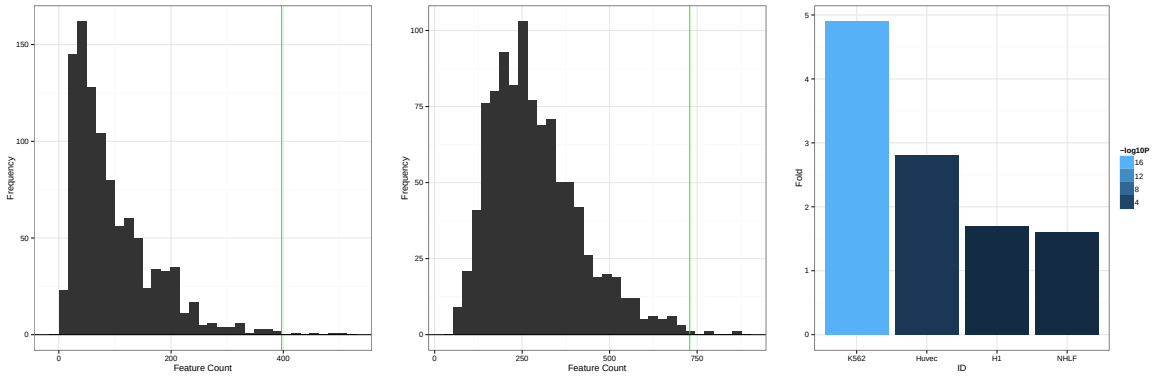

Figure 2.3: **Replicated distant miRQTL are enriched for ENCODE annotations** We created n=1000 bins of minor allele frequency (MAF) matched SNPs to the 67 replicated distant miRQTL. For each bin, we counted the number of genomic features and compared the result to the that found with the replicating miRQTL. A) Replicating distant miRQTL were more likely to be functionally annotated as protein binding (ChIP-seq) B) Replicating distant miRQTL were more likely to be functionally annotated to regions containing open chromatin (DNAse-seq, FAIRE) C) Using HaploReg we found that miRQTL were enriched for being located within myeloid derived K562 cell specific histone enhancers regions.

dance. Evidence that distant QTL act through regulation of local transcripts has previously been provided [86]. To this end, we performed eQTL mapping restricting to the 67 replicated miRQTL SNPs using the FHS dataset. We identified 493 local (cis-chromosome) eQTL associations (FDR < 0.05) for 35 of the 67 tested SNPs which corresponded to 18 transcripts (Table 2.1, Supplementary Table 3). These 35 eQTL corresponded almost entirely to miRQTL for hsa-miR-185-5p (490 of 493 SNP-transcript pairs). Furthermore 16 of 18 eQTL-associated transcripts were expressed on chromosome 17.

Table 2.1: eQTL mapping with replicating miRQTL

SNP	Gene	tstat_Gene	pvalue_Gene	FDR_Gene	tstat_Gene_miR	pvalue_Gene_miR	FDR_Gene_miR	miR
rs4795456	LRRC75A-AS1	4.420	0.0000100641514577841	0.005211368163061	3.10380	0.00192	0.06710	hsa.miR.185.5p
rs4795456	HEXIM1	-4.286	0.0000185167843126471	0.008525565703053	-3.04415	0.00235	0.07548	hsa.miR.185.5p
rs9310736	TP63	3.825	0.0001321600133325700	0.043202152207118	-1.97680	0.04812	0.35913	hsa.miR.183.3p
rs4795456	FAM104A	-4.274	0.0000195830900495312	0.008902560872496	-1.93233	0.05338	0.37542	hsa.miR.185.5p
rs4795456	MYL4	8.011	0.0000000000000014149	0.000000000005549	1.81570	0.06948	0.41946	hsa.miR.185.5p
rs4795456	SLC4A1	-4.257	0.0000211165296186298	0.009416191504779	1.68388	0.09227	0.47065	hsa.miR.185.5p
rs901975	SLC25A39	-4.576	0.0000048656277956570	0.002831697450073	1.41892	0.15599	0.57422	hsa.miR.185.5p
rs4795456	FAM222B	-5.254	0.0000001549021477453	0.000125309889555	-0.93425	0.35022	0.75567	hsa.miR.185.5p
rs4795456	AP2B1	-4.474	0.0000078578228990065	0.004162436123783	0.91552	0.35997	0.76210	hsa.miR.185.5p
rs901975	TIAF1	-4.006	0.0000627826551706548	0.023141190836892	-0.62687	0.53077	0.85449	hsa.miR.185.5p
rs901975	PCTP	-4.445	0.0000089802018294138	0.004719684927051	-0.50888	0.61086	0.88814	hsa.miR.185.5p
rs13019832	RPIA	4.738	0.0000022247277802850	0.001382008581514	0.47564	0.63435	0.89714	hsa.miR.192.5p
rs13019832	TNS1	-5.453	0.0000000519529918716	0.000048738549220	-0.43487	0.66368	0.90785	hsa.miR.192.5p
rs901975	GID4	4.486	0.0000074402112806278	0.003952577400346	0.31784	0.75062	0.93655	hsa.miR.185.5p
rs4795456	UBE2O	-4.724	0.0000023759374296360	0.001472838519197	-0.31176	0.75524	0.93799	hsa.miR.185.5p
rs4795456	MAP2K3	-8.951	0.0000000000000000005	0.000000000000003	0.22874	0.81908	0.95645	hsa.miR.185.5p
rs4795456	ALOX15	3.876	0.0001077789112821080	0.036618719525305	0.08434	0.93279	0.98521	hsa.miR.185.5p
rs4795456	RAB34	-4.054	0.0000510459975453588	0.019655201379125	0.02469	0.98031	0.99578	hsa.miR.185.5p

We reasoned that a microRNA distantly regulated through local transcript eQTL should show evidence of association with the corresponding local mRNA transcript. Therefore we tested the association between the 18 local eQTL-associated mRNA transcripts and the three microRNA associated with the 35 miRQTL. We identified five significantly associated transcript-microRNA pairs (FDR < 0.5), corresponding between transcripts, *SLC4A1*, *MYL4*, *FAM104A*, *HEXIM1*, *TP63*, *LRRC75A-AS1* and microRNA hsa-miR-185-5p and hsa-miR-183-3p (Supplementary Table 3). *TP63* was the only transcript associated with hsa-miR-183-3p.

To strengthen the support for the hypothesis that distant miRQTL hsa-miR-185-5p, hsa-miR-183-3p and hsa-miR-192-5p are acting through regulation of local transcripts, we performed mediation analysis using the 35 loci, 18 transcripts and 3 microRNA previously identified. We found that four transcripts *MYL4*, *HEXIM1*, *TNS1* and *AP2B1* reported Sobel p-values < 0.05 (FDR < 0.03) supporting the hypothesis for mRNA transcript mediated microRNA abundance (Table 2.2, Supplementary Table 3). Interestingly, we found that *HEXIM1* showed a negative direction of effect with hsa-mir-185-5p expression suggesting the possibility of a novel function for *HEXIM1* as a microRNA repressor. Over all tests, *HEXIM1* presented the strongest evidence of being a locally regulated mRNA transcript that influences microRNA abundance as it was the only transcript that showed evidence of transcript-microRNA association ($p < 0.05$) and mediation ($p < 0.05$).

Discussion

MicroRNAs are important, evolutionarily conserved, regulators of gene expression and translation. Improving our understanding of genetic mechanisms impacting on microRNA abundance is key to elucidating cellular and physiological processes. Previous studies have demonstrated that common genetic variants influence differences in microRNA levels between individuals, however, they were primarily focused on local regulation. Our study aimed to

Table 2.2: Mediation analysis of lead SNP-transcript pairs

SNP	miR	Gene	Sobel_Z	Sobel_P
rs4795456	hsa.miR.185.5p	MYL4	2.217	0.027
rs4795456	hsa.miR.185.5p	HEXIM1	1.985	0.047
rs13019832	hsa.miR.192.5p	TNS1	1.976	0.048
rs4795456	hsa.miR.185.5p	AP2B1	-1.909	0.056
rs9310736	hsa.miR.183.3p	TP63	-1.600	0.110
rs901975	hsa.miR.185.5p	GID4	-1.578	0.115
rs4795456	hsa.miR.185.5p	SLC4A1	-1.463	0.143
rs4795456	hsa.miR.185.5p	LRRC75A-AS1	1.315	0.189
rs901975	hsa.miR.185.5p	SLC25A39	-1.207	0.227
rs4795456	hsa.miR.185.5p	MAP2K3	-1.083	0.279
rs13019832	hsa.miR.192.5p	RPIA	-1.039	0.299
rs4795456	hsa.miR.185.5p	FAM104A	0.917	0.359
rs4795456	hsa.miR.185.5p	RAB34	0.736	0.462
rs4795456	hsa.miR.185.5p	UBE2O	0.731	0.465
rs901975	hsa.miR.185.5p	TIAF1	0.716	0.474
rs901975	hsa.miR.185.5p	PCTP	0.390	0.696
rs4795456	hsa.miR.185.5p	ALOX15	-0.352	0.725
rs4795456	hsa.miR.185.5p	FAM222B	0.131	0.896

elucidate putative genetic mechanisms governing distant regulatory biology of microRNA expression. We systematically examined the distant regulatory landscape of microRNAs by expanding on the largest miRQTL study to date and taking advantage of recent methods that effectively correct for unmeasured confounding factors.

Our *in silico* analyses suggest novel regulatory mechanisms for microRNA regulation. We identified putative local regulators of mRNA gene expression which subsequently may regulate distant microRNA. *HEXIM1* provided the strongest evidence for being such a regulator as it showed association between hsa-mir-185-5p ($p < 0.05$) and mediation analysis confirmed it as a putative mediator ($p < 0.05$). HEXIM1 is a transcriptional regulator that acts to inhibit RNA polymerase II transcript elongation by sequestering the P-TEFb complex[16]. Importantly HEXIM1 is a double strand-RNA binding protein that canonically works in concert with 7SK snRNA[63]. Experimental evidence suggests that its regulatory roles may be diverse, possibly regulating transcriptional activity of NF-kappa-B, ESR1, NR3C1 and

CIITA[54, 94, 124, 76, 130, 77, 83]. As HEXIM1 is a key regulator of RNA polymerase II, a protein shown to elongate microRNA transcripts[61], it may thus be a regulator of hsa-mir-185 through inhibition of transcriptional machinery. Alternatively, Li et al. tested whether HEXIM1 was capable of binding microRNA *in vitro*[63]. They found that the microRNA they tested (miR-16-1) could be detected in HEXIM1 immunoprecipitates. The authors did not detect the pre-mir-16-1 version in their immunoprecipitate which lead them to speculate that HEXIM1 may bind double stranded microRNA intermediates. Intermediate double stranded microRNA bound to HEXIM1 could reduce the availability of downstream mature microRNA products. Taken together, our analytical results suggest a negative regulatory role for HEXIM1 in modulating microRNA levels, either through its effects on transcription, or through its ability to bind short RNA molecules.

Albeit with weaker analytical support (transcript-miR $p < 0.05$, mediation $p = 0.1$), we observed an intriguing association between hsa-mir-183-3p and *TP63*. *TP63* encodes the protein P63, which has been shown to have complex regulatory interactions with microRNA, as some microRNA regulate *TP63* and P63 directly regulates microRNA processing components Dicer and DGCR8. The existence of evidence directly implicating TP63 in microRNA regulation supports our in silico analyses pointing to its potential role in modulating hsa-miR-183 abundance.

In addition, the 35 miRQTL distantly associated with microRNA are entirely contained either within or near the gene *FAM222B* and *ERAL1*. *ERAL1* itself is an RNA binding GTPase that has been shown to bind a 33 nucleotide region of 12S mitochondrial rRNA (12S mt-rRNA)[22]. X-ray crystallography of the ERAL1 *E. coli* ortholog ERA revealed a KH-domain as the functional RNA binding region[107, 106]. Single strand binding KH-type proteins such as KHSRP and IGF2BP1 have been shown to play a role in microRNA biogenesis and a variety of microRNA mediated cellular processes[108]. While the structure of ERAL1 in humans remains unresolved, it may share homology with human microRNA

binding proteins which could influence microRNA processing and abundance.

In closing, our findings support the idea that comprehensive miRQTL mapping can provide valuable insight into genetic mechanisms underlying distant regulation of miRQTL, thereby impacting on our understanding of gene regulation and associated phenotypes. This finding has the ability to expand and enrich our understanding of the mechanisms governing microRNA and mRNA transcript abundance.

Methods and Analysis

Data Acquisition and microRNA expression profiling and genotyping

MiRNA expression levels were downloaded from dbGaP ($phs000007.v21.p8$, $phs000007.v21.p8$). Prior to data posting on dbGaP, microRNA expression profiling and genotyping were as described in the methods section of Huan et al. 2015[46]. Therefore we briefly describe the enrollment, genotyping and microRNA profiling. This study leverages the offspring cohort (n=2,272) and the third generation cohort (n=3057) of the original FHS cohort which began enrollment in 1948. Whole blood was collected from study participants using the PAXgene Blood RNA tubes and RNA was extracted using a PAXgene Blood RNA extraction kit. A 2100 Bioanalyzer Instrument was used to evaluate RNA quality. The RNA samples were then converted to complementary DNA (cDNA) with a TaqMan microRNA Reverse Transcription Kit MegaPlex Human RT Primer Pool Av2.1 and Pool Bv3.0. TaqMan PreAmp Master Mix and PreAmp Primers, Human Pool A v2.1 and Pool B v3.0 kits were used for PreAmplification. Subsequently the PreAmplified cDNA samples were run on a Fluidigm BioMark System with TaqMan microRNA Assays. The Affymetrix GeneChip Human Mapping 500K Array and the 50K Human Gene Focused Panel were used to obtain results for 9,274 unique FHS participants, including 1,529 original cohorts, 3,753 offspring, 99 offspring spouses, and 3,893 third generation participants. Genotypes were called using the BRLMM

algorithm and imputed to HapMap 2 using MACH.

Reproducibility

Our analysis aims to be fully reproducible. All code used for quality control and analysis can be found on github at *https://github.com/hakyimlab/ERAL1* which will allow for detailed evaluation and reproduction. Full data sets will require dbGaP access.

microRNA Analysis Quality Control

To increase power to detect distant regulatory associations we removed microRNA that were expressed in less than 50% of the individuals (n=5434). This resulted in 202, expressed in over 50% of the individuals (Supplementary Figure 2). For the raw cycle threshold (CT) matrix, we imputed data which were missing for CT values greater than 27 by sampling uniformly real numbers between 27 and 35. We then converted CT values to interpretable expression value by subtracting 35 from all CT (Supplementary Figure 3). We then log2 transformed the data. The imputed expression matrix was used for the calculation of PEER factors exclusively. The statistical models using microRNA expression were all fit with missing=NA for missing expression values.

It is well established that expression profiling captures sources of variation other than genotype[99]. For example technical sources may contribute to variation in microRNA concentration values among individuals. Examples of sources include (i) RNA isolation batch, (ii) RNA quality, (iii) RNA concentration, and (iv) 261/280 ratio. Furthermore there may be unknown and potentially unwanted factors captured in the expression data. Accounting for these factors has been shown to improve detection of genotype-expression associations[99]. In contrast to Huan et al, we adjusted for unwanted variation using PEER factors, which has been shown to increase the power to detect eQTL, a property that logically extends to miRQTL. We used 40 factors consistent with the large scale distant-eQTL mapping study

by Westra et al. [119].

Genotype Analysis Quality Control

Starting with imputed genotypes obtained from dbGaP, we removed SNPs that were missing in more than 5% of the individuals and excluded individuals missing over 5% of their SNPs using plink[87]. Additionally using plink, we excluded SNPs with HWE p-values less than 0.05 and removed SNPs with a minor allele frequency (MAF) of less than 0.01. This resulted in 2,543,887 SNPs in 6810 individuals. The details of this quality control process can be found in the repository at *data/matrixeqtl_related/FramImpQC1.log*

The full association model was fitted as:

$$\mathrm{miR}_k \sim \mathrm{SNP}_j + \mathrm{PEER}_1^{\mathrm{miR}} + \cdots + \mathrm{PEER}_n^{\mathrm{miR}} \tag{2.1}$$

where k is the index for microRNA, j is the index for SNP, n is the number of PEER factors used. We observed a well controlled genomic inflation factor λ of 1.02 and quantile-quantile (QQ) plot (Figure 1A). Finally because of the burden of fitting linear-mixed models in eQTL studies, we used faster linear regression for discovery and confirmed the top associations remained unchanged using *lmekin* in R as per the methods of Huan et al. 2015.

We restricted the output of Matrix eQTL to results with false discovery rate (FDR) < 0.05 resulting in 1,665 SNP-microRNA pairs. We then annotated the results with NCBI SNP and gene information using the R packages *ncbi2r* and *mirbase.db*. We annotated a SNP-microRNA association as local if the SNP was located on the same chromosome as the miRBase annotated transcription start site (TSS). Otherwise a SNP-microRNA association was annotated as distant.

Replication

We replicated the results in the Geuvadis dataset using data obtained from *http://www.ebi.ac.uk/Tools/gEUVADIS-das/*. Specifically we used the file GD480.MirnaQuantCount.txt as it allowed flexibility in correcting for unwanted variation. The intersection of genotype, microRNA and mRNA expression data resulted in a dataset comprised of 331 individuals. We matched by dosage allele and confirmed that MAF between datasets was within 0.15 of each other (Supplementary Figure 4). We found that including 10 PEER factors maximized the number of replicating miRQTL (Supplementary Figure 5). Replication consisted of fitting the following models:

$$\text{miR}_i \sim \text{SNP}_j + \text{PEER}_1^{\text{miR}} + \cdots + \text{PEER}_n^{\text{miR}} \tag{2.2}$$

where i indexes microRNA (FDR < 0.05 in FHS), k indexes the SNP (FDR < 0.05 in FHS) and n is the number of PEER factors used.

eQTL Analysis

We performed eQTL mapping in the FHS dataset using the 67 replicating SNPs previously identified and 22,011 transcripts measured on the Affymetrix Human Exon array. We included 20 PEER factors calculated from the full gene expression matrix. Therefore we fit

$$\text{transcript}_i \sim \text{SNP}_j + \text{PEER}_1^{\text{transcript}} + \cdots + \text{PEER}_n^{\text{transcript}} \tag{2.3}$$

where i indexes the transcript, j indexes the SNP and n is the number of PEER factors used. We tested the association between transcript level and microRNA level by

$$\text{miR}_i \sim \text{transcript}_j + \text{PEER}_1^{\text{transcript}} + \cdots + \text{PEER}_n^{\text{transcript}} + \text{PEER}_1^{\text{miR}} + \cdots + \text{PEER}_m^{\text{miR}} \tag{2.4}$$

where i indexes the miR, j indexes the transcript, n is number of transcript-PEER factors used, and m is the number of microRNA-PEER factors used.

Mediation Analysis

Mediation analysis was conducted using the function *mediation.test* found in the github cran repository *https://github.com/cran/bstats/tree/master/R*. The application of mediation analysis in eQTL studies has been investigated by Pierce et al. 2014[86]. Briefly, within the context of eQTL studies, one may regress the dependent distant associated probe (in this case microRNA) onto the independent variable SNP genotype plus mediating variable mRNA probe:

$$\mathrm{Y_{miR}} = \beta + \beta \mathrm{X_{SNP}} + \beta_1 \mathrm{X_{ERAL1}} + \epsilon \tag{2.5}$$

where ϵ is the error term. We also must regress the mediating local associated probe (in this case mRNA probe for *ERAL1*) onto the independent variable SNP genotype:

$$\mathrm{Y_{ERAL1}} \sim \beta + \beta_2 \mathrm{X_{SNP}} + \epsilon \tag{2.6}$$

Our t-statistic:

$$\mathrm{SE} = \sqrt{\beta_1^2 \sigma_{\beta_2}^2 + \beta_2^2 \sigma_{\beta_1}^2} \tag{2.7}$$

$$\mathrm{t} = \beta_1 \beta_2 / SE \tag{2.8}$$

is finally compared to the normal distribution to generate sobel p-values. In addition we wanted to incorporate the covariates used in equations 2, 3 and 4. Therefore in practice we fit:

$$\text{COVAR} = \text{PEER}_1^{\text{miR}} + \cdots + \text{PEER}_n^{\text{miR}} + \text{PEER}_1^{\text{transcript}} + \cdots + \text{PEER}_m^{\text{transcript}} \quad (2.9)$$

$$\text{miR-589} \sim \text{EXON1}_{\text{ERAL1}} + \text{rs4795456} + \text{COVAR} \quad (2.10)$$

$$\text{EXON1}_{\text{ERAL1}} \sim \text{rs4795456} + \text{COVAR} \quad (2.11)$$

where n is the number of microRNA PEER factors used and m is the number of transcript PEER factors used. COVAR are the covariates of equations 2, 3 and 4.

Supplementary Figures

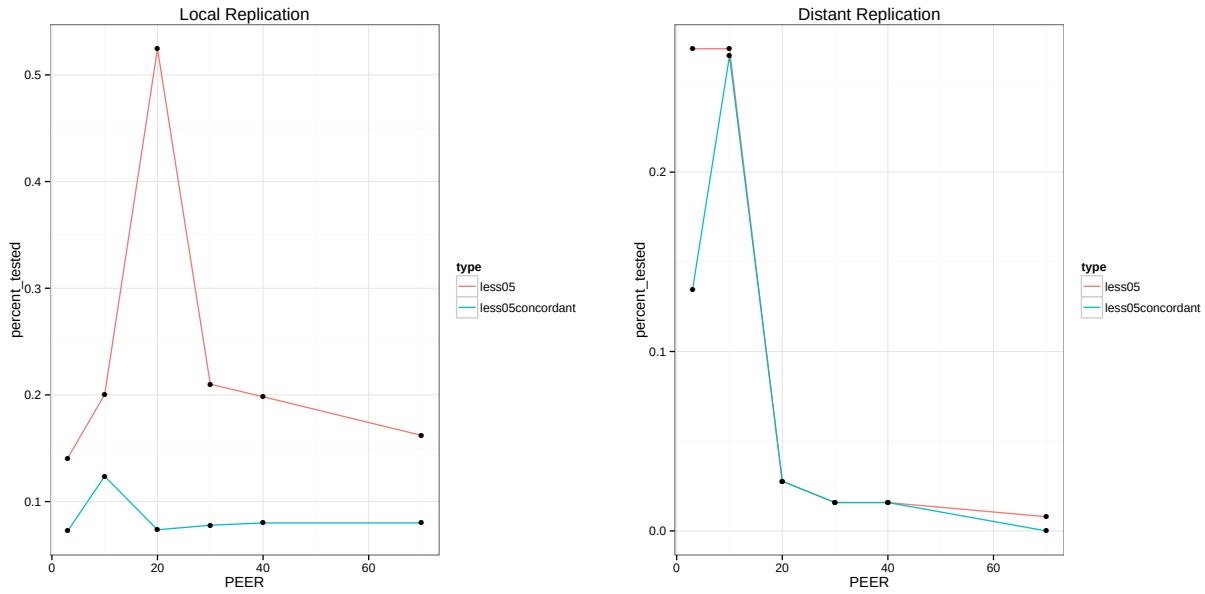

Figure 2.4: **Supplementary Figure 1 A-B** Concordance plot between FHS fitted with 40 PEER factors and Geuvadis datasets over a range of PEER adjusted comparisons in Geuvadis. Here we compare the percentage of SNPs that have either p-values < 0.05 or p-values < 0.05 with concordant direction of effect. A) We observe 55% of tested loci with p-values < 0.05 when fit with 20 PEER factors however the percentage concordant remains low at 0.05%. B) We see that for distant miRQTL, at 10 PEER factors fitted in Geuvadis, we achieve concordant direction of effect which is maintained up to 40 PEER factors. At 70 PEER factors we begin to see diverging concordance.

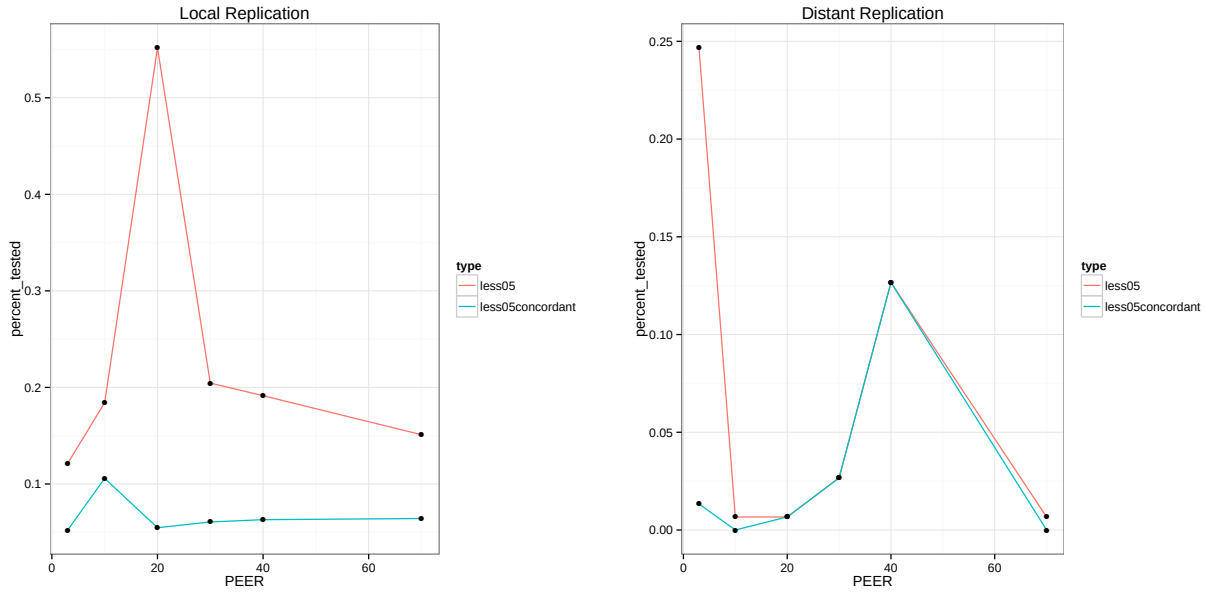

Figure 2.5: **Supplementary Figure 1 C-D** Concordance plot between FHS fitted with 20 PEER factors and Geuvadis datasets over a range of PEER adjusted comparisons in Geuvadis. Here we compare the percentage of SNPs that have either p-values < 0.05 or p-values < 0.05 with concordant direction of effect. C) We see similar results to parts A where 20 PEER factors fitted in Geuvadis maximizes the percentage tested with 52% of local tests showing p-values < 0.05 yet only 0.07% are concordant. D) We see that at 40 PEER factors fitted in Geuvadis, we achieve concordant direction of effect with direction maintained up to 70 PEER factors.

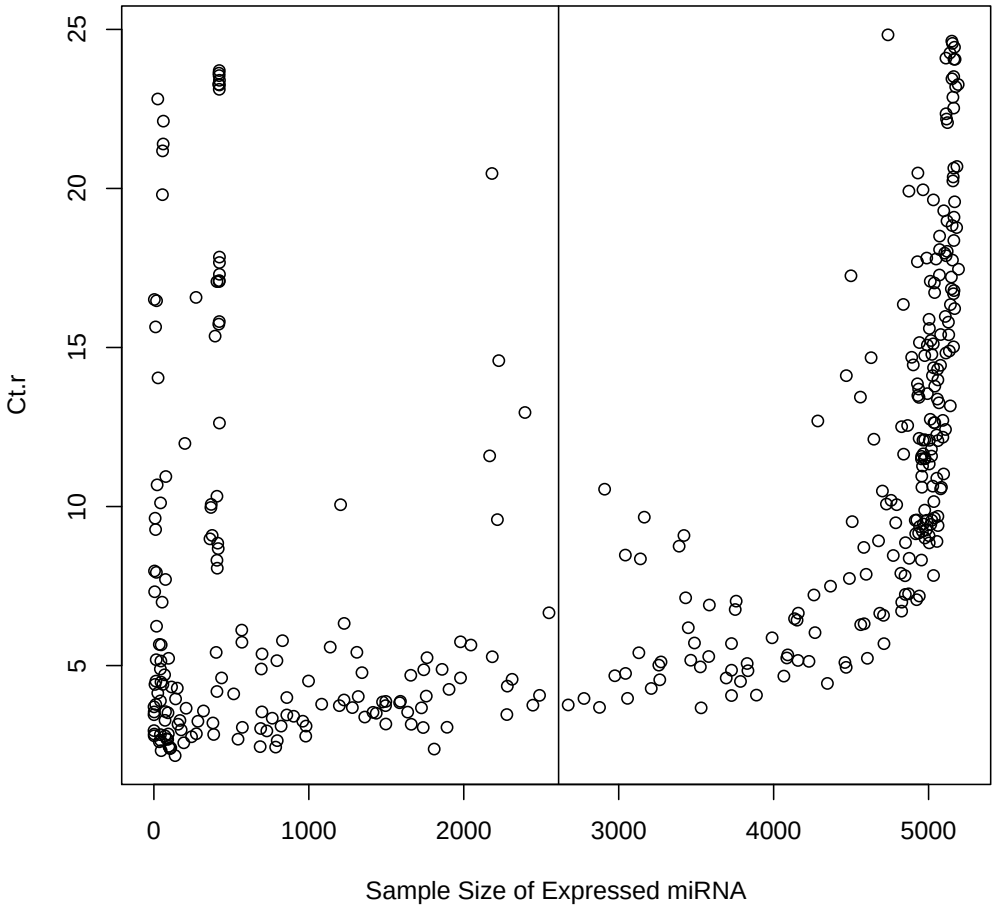

Figure 2.6: **Supplementary Figure 2** Distribution of sample size of the expressed microRNA and the mean CT expression. Missing values were uniformly imputed between values of 27-35. The raw CT values were subtracted from cycle threshold 35. The vertical line represents the 50% cutoff used for miRQTL mapping.

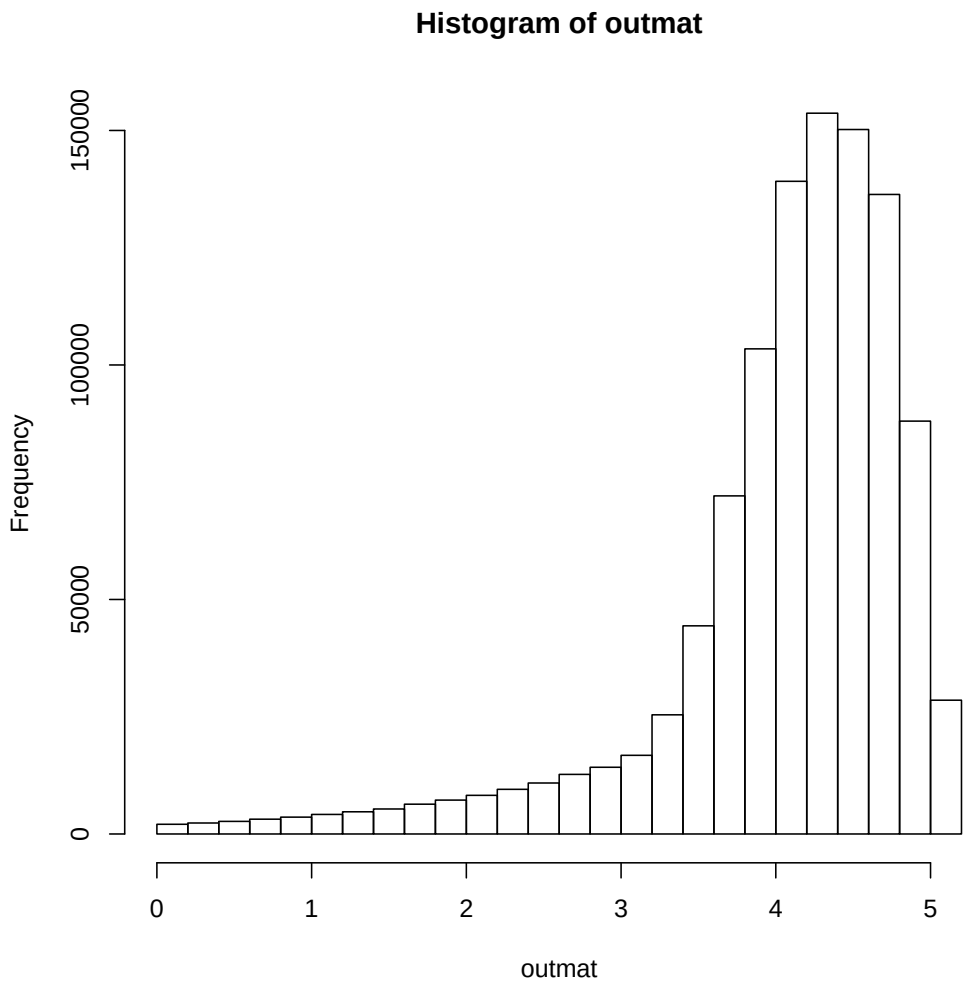

Figure 2.7: **Supplementary Figure 3** Distribution of log2 transformed mean CT levels in the FHS dataset. The data were imputed for CT values greater than 27 and subsequently all values were subtracted from 35. Values <= log2(8) were imputed in this figure.

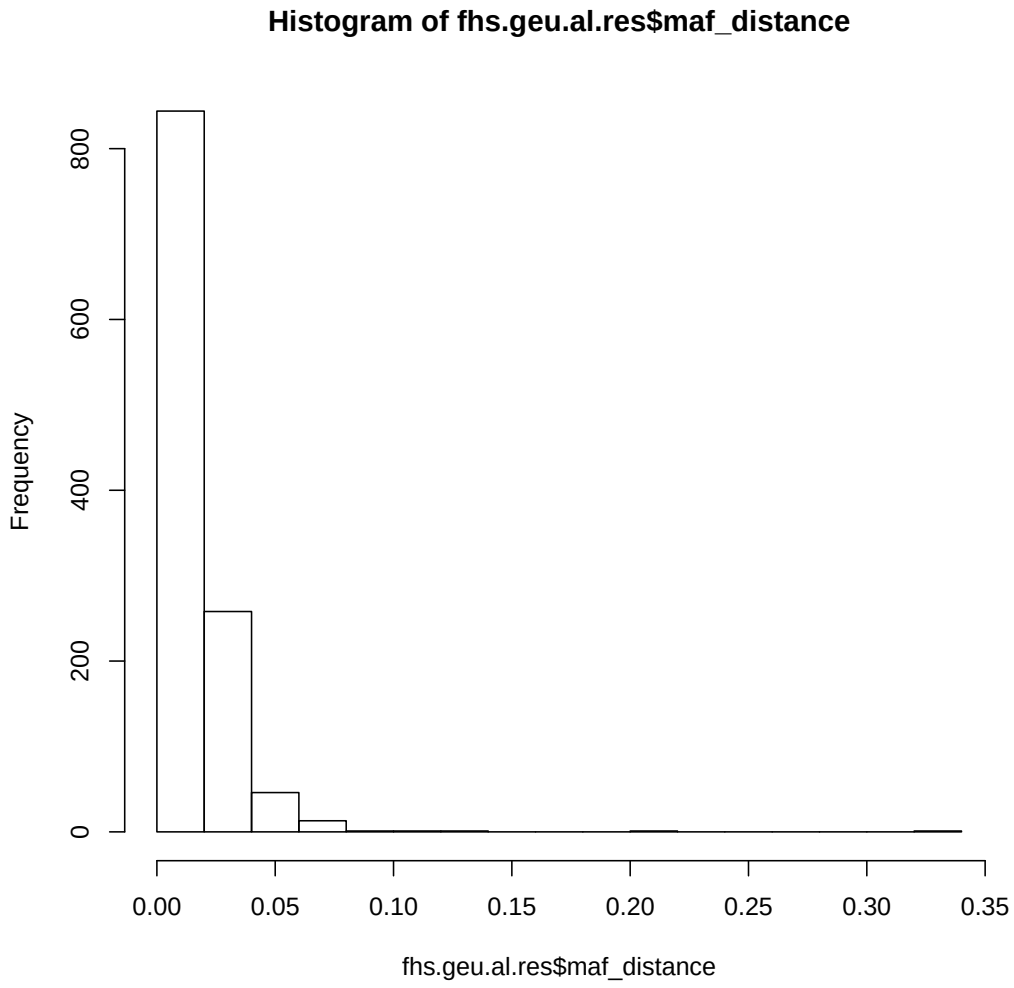

Figure 2.8: **Supplementary Figure 4** Difference in minor allele frequence (MAF) between Framingham Heart Study and Geuvadis alleles. Prior to replication we removed SNPs which had differences greater than 15%.

CHAPTER 3

POLY-OMIC PREDICTION OF COMPLEX TRAITS: OMICKRIGING

Abstract

Clinically relevant prediction of complex traits is an ultimate goal of precision medicine. Recently genome wide association (GWA) studies have uncovered thousands of trait-associated loci. However trait-associated loci were shown to collectively explain only a small portion of phenotypic variance, thereby limiting their utility in a clinical setting. To address this limitation, we developed a fast, scalable and highly versatile approach to whole-genome prediction of complex traits. Our approach uses a method called Kriging, commonly used in geostatistics and machine learning. We extend Kriging to the Omic setting to present OmicKriging. Intuitively, OmicKriging translates genetic similarity between individuals into phenotypic similarity. We show that OmicKriging is capable of integrating multiple sources of systems level data to predict unobserved phenotypes. In addition we show that OmicKriging flexibly addresses the genetic architecture of complex traits as it leverages previously identified loci identified through GWAS. Using seven disease datasets from the Wellcome Trust Case Control Consortium (WTCCC), we show that OmicKriging allows simple integration of sparse and highly polygenic components yielding comparable performance at a fraction of the computing time of a recently published Bayesian sparse linear mixed model method. Using a cellular growth phenotype, we show that integrating mRNA and microRNA expression data substantially increases performance over either dataset alone. We provide an R package to implement OmicKriging (*http://www.scandb.org/newinterface/tools/OmicKriging.html*).

Introduction

Clinically relevant prediction of complex traits is an ultimate goal of precision medicine. Recently genome-wide association (GWA) studies were undertaken to gain novel biological insight and yield clinically actionable trait-associated loci. The result of GWAS have been staggeringly successful having identified 19,603 genetic loci to date. However the small effect sizes of the vast majority of trait-associated loci has limited their clinical utility[17]. Recently, the use of polygenic models has shown that for complex traits, a substantially greater proportion of phenotypic variance can be explained over the use of single variant tests[88, 128]. A notable example is height where approximately 45% of the phenotypic variance can be explained using a mixed linear modeling approach (GCTA) that simultaneously considers all 300K common SNPs genotyped [128, 129]. GCTA circumvents limitations of GWAS by not relying on the selection of loci when estimating the proprortion of variance explained. This approach has been considered more appropriate in the context of highly polygenic traits[71, 80].

To gain an intuition for OmicKring, we consider an analogous application from the field of geospatial statistics. Kriging estimates the measurement (rainfall) at an unobserved location through interpolation of measured locations[12, 101, 48]. Here we assume that the rainfall at close locations will be more similar to the unmeasured location than rainfall at distant locations. Kriging weighs each observed measurement by the distance to the location of the unobserved quantity to make a prediction. Analogously, we consider these locations as individuals, measurements as phenotypes, and distance as genetic similarity(Figure 3.1). Close ties between genetic distance and geographic distance have been demonstrated in studies of human population structure. For example, in the analysis of genome-wide genotype data from 3000 Europeans, the first two principal components generated a graphic resembling the geographic map of Europe [79]. A diagram of the analogy between geostatistical kriging and complex trait prediction is shown in Figure 3.1.

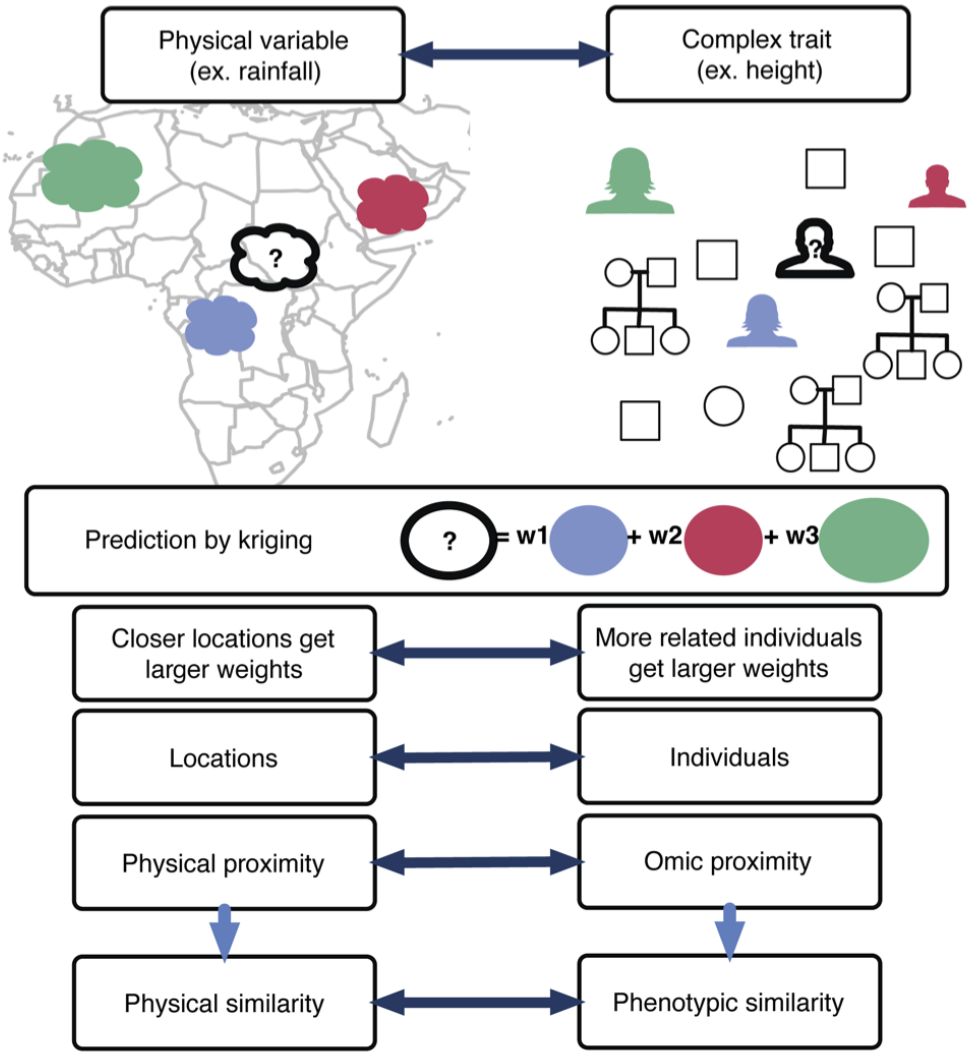

Figure 3.1: **Kriging and whole-genome prediction connection.** This figure shows the analogous relationships between components of the kriging method used in geostatistics and whole-genome prediction. The prediction at an unobserved location (?) is computed as a weighted average of the variable at observed locations. The weights are functions of the correlation between the rainfall at the new location and the rainfalls at the observed locations. The closer the distance between each observed location and the new location, the higher the weight. In complex trait prediction, locations correspond to individuals, physical proximity corresponds to genetic relatedness. The correlation between two locations or individuals is the key component of this method. In animal breeding approaches, the genetic relatedness matrix or kinship matrix is used. In OmicKriging, a genetic relatedness matrix, a gene expression similarity matrix, or any combination of available high-throughput data similarity measures can be tested for complex trait prediction performance.

Whole-genome prediction was initially introduced by Meuwissen et al. [75] who proposed to predict unobserved phenotypes using out-of-sample-estimated regression coefficients which were linearly combined with SNP genotype: $\sum_{l=1} X_{il}\hat{\beta}_l$, where X_{il} is the genotype of individual i for marker l and $\hat{\beta}_l$ is the estimated effect size of marker l. This approach is known as the polygenic score method. The polygenic score method saw an early application to human disease risk using logistic regression effect sizes for SNPs at or below an arbitrary threshold. The approach was successful at predicting a significant proportion of the risk of schizophrenia and bipolar disorder when generalized to test cohorts[88]. Purcell et al. trained logistic regression models using a large schizophrenia GWAS cohort and used test cohorts which included additional schizophrenia and bipolar disorder patients. A training-test set approach that used the polygenic score method was similarly tested in the Wellcome Trust Case Control Consortium (WTCCC) seven disease cohort[25].

There exist several additional whole-genome prediction (WGP) methods, which were reviewed by de los Campos et al. [17] and Abraham et al. [1]. These methods include penalized estimation models such as Least Absolute Shrinkage and Selection Operator (LASSO) [105], Ridge Regression [44], and Elastic Net [136, 1]. These approaches cleverly exploit geometric properties of their models to penalize the coefficients used to estimate an unknown fixed vector $\beta \in \mathbb{R}^p$. These approaches are particularly valuable when working with underdetermined systems of equations as is the case with GWAS level data[17]. For example, Vazquez et al. applied a Bayesian LASSO to structure the prior density of marker effects in their Bayesian regression WGP model of skin cancer risk [110]. Their best prediction model, which included 41K genome-wide SNPs, had an area under the receiver operating characteristic curve (AUC) of 0.635, 18.9% higher than that of the baseline model, which just included non-genetic covariates [110].

The notion of translating genetic similarity into phenotypic similarity which is used to predict phenotypes dates back to Fisher [27] and Wright [127]. These ideas were formalized as

best linear unbiased prediction (BLUP) approaches based on multivariate normal processes by Henderson, Goldberger and others [33, 40, 41]. When these methods were first developed, high-throughput genotyping was not available therefore the methods used pedigree based similarity matrices. Recently with the advent of affordable high throughput genotyping technologies, several authors have used similarity measures computed using larger numbers of genetic markers [18, 35, 51, 69, 71, 80, 109, 128]. These approaches are typically referred to as G-BLUP for genomic-BLUP. G-BLUP differs from the polygenic score approach by its ability to fit all markers of the genome simultaneously. G-BLUP performs this by using only a matrix of genetic relatedness (GRM) calculated from SNP genotype data. Recently, the results of WTCCC analysis reveal that the genetic architecture of complex traits can be highly variable with some traits owing much of their influence to a small number of loci of modest effect and others to many loci of small effect. This motivated the development of genetic architecture-flexible approaches such as a the model developed by Zhou et al which combines G-BLUP with a sparse regression model that allows for a small proportion of markers with large effect sizes and estimates the most likely model for a particular phenotype from the data [133].

Furthermore, Kriging and BLUP have well established use in animal breeding and quantitative genetics fields [37, 90]. Kriging in genomic prediction has been previously used, but it was restricted to simulation studies of genetic similarities [81]. Based on whole genome simulations, Ober et al. reported that using Matérn functions to scale the genetic relatedness measure works better than standard measures of relatedness in the presence of dominance and epistatic effects [81].This suggests that alternate approaches to distance matrix calculation may yield domain specific prediction improvements.

Here we extend Kriging to the omic setting for use with multiple sources of high-throughput systems level data. In addition we show that OmicKriging flexibly integrates external information to improve overall phenotype prediction. For example, previously iden-

tified trait-associated loci can be given greater weight to improve prediction performance. This ability differentiates OmicKriging from standard Kriging/BLUP methods in that it is not necessarily tied to an additive genetic/genomic model. Rather than using maximum likelihood estimation methods, we search for optimal hyper-parameters which maximize cross-validated prediction performance. In this sense, our approach is closely related to the semi-parametric models using reproducing kernel Hilbert space (RKHS) regression proposed by Gianola et al. [32] and de los Campos et al. [19] for WGP. To our knowledge, our method is the first to integrate multiple omics data using these semi-parametric methods based on similarity measures. Importantly we show that OmicKriging is a fast and scalable whole-genome prediction method for high-throughput omic technologies.

Results

To test the prediction performance of OmicKriging we first calculated inter-sample similarity matrices for each respective high-throughput source. These matrices included the genetic relationship matrix (GRM) calculated from SNPs and a gene expression correlation matrix (GXM) calculated from gene expression data. We then predicted unknown phenotypes by calculating the weighted average of the training set individuals' phenotypes. In the case of SNP data, the weights were comprised of the GRM and pairwise genotype similarity of the unknown individual with the genotypes of those with observed phenotypes. When using a single omic component, we tested matrix weights between 0 and 1 (i.e. 0.1, 0.2, 0.3, ..., 1 for the omic component and 1-weight for the environmental component) to find the matrix weight that produced optimal prediction. When two omic components (e.g. GRM and GXM) were combined, we performed a grid search to find the optimal prediction matrix weights θ_1 and θ_2, such that $\theta_1 + \theta_2 \leq 1$ (1-θ_1-θ_2 for the environmental component). We found that the optimal matrix weights for each omic component depended on the genetic architecture of the trait. A schematic of our procedure to find the optimal composite similarity matrix

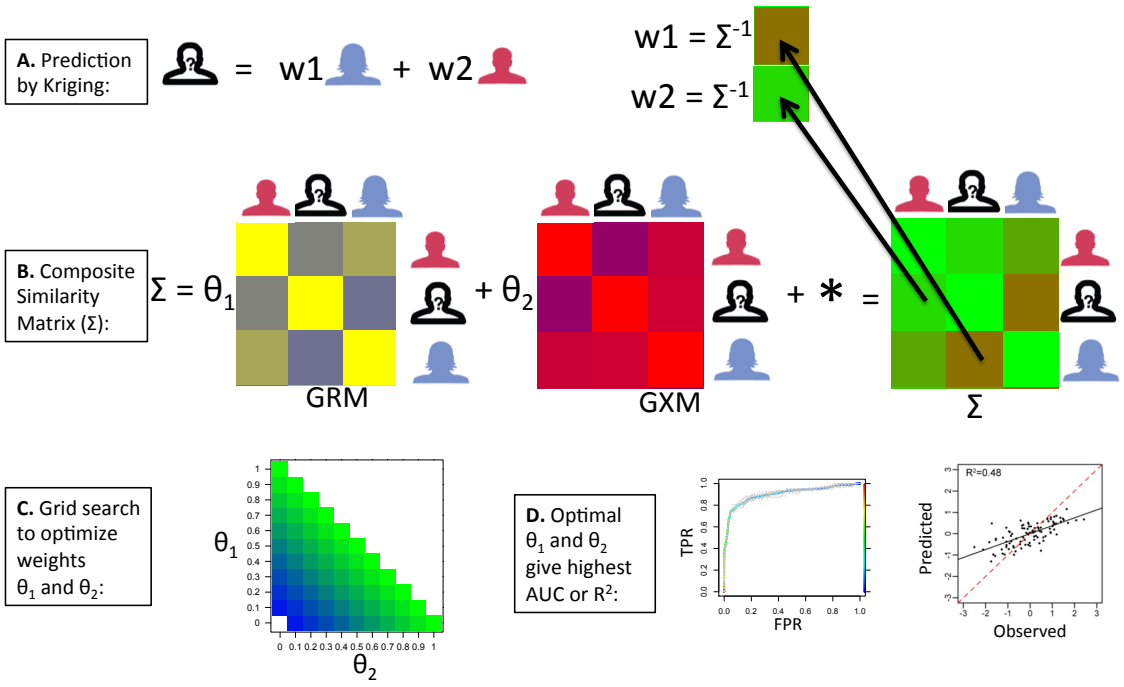

Figure 3.2: **OmicKriging data integration and weighting.** (A) The individual weights, depicted as w1 and w2, in the Kriging method are given by the product of the composite similarity matrix Σ and the correlation of omic data between the individual of unknown phenotype (?) and the individuals of known phenotype. (B) The composite similarity matrix Σ integrates different omic correlation matrices such as a genetic relationship matrix (GRM) derived from SNPs and a gene expression correlation matrix (GXM) derived from gene expression levels in this example. Σ also includes an environmental component, i.e. noise term (*). (C) In OmicKriging, we optimize the matrix weights, θ_1 and θ_2, by testing the θ_i values of the grid space depicted in color. (D) The optimal matrix weights θ_i give the highest values of AUC for binary traits and R^2 for quantitative traits.

is shown in Figure 3.2.

Our implementation of OmicKriging allows for parallel computation of each k of k-fold cross-validation. Therefore we used a 16-fold cross-validation approach reflective of our compute resources, with individuals assigned to the 16 subsets at random and repeating the procedure 500 time to assess the sampling variability of the prediction performance (see Methods). For quantitative traits, we computed the coefficient of determination R^2 (equivalent to the square of the correlation) between the predicted and true values of the

phenotype. We assessed prediction performance and for case/control traits by computed the area under the receiver operating characteristic curve (AUC).

Cellular phenotype applications

To test the prediction performance of OmicKriging when combining multiple sources of omic data, we used the intrinsic growth rate (iGrowth) phenotype derived from multiple proliferation measurements in the commonly used HapMap lymphoblastoid cell lines (LCLs) [47]. We felt this was an appropriate phenotype to demonstrate the performance of OmicKriging as it has been shown that genes associated with proliferation are strong prognostic factors in several types of cancers [98, 15, 13, 91] and such genes are differentially expressed in most cancer tissues [122, 92, 89].

Therefore we used iGrowth values from 99 LCLs from the HapMap CEU (Northern and Western European ancestry from Utah) and YRI (Yoruba from Ibadan, Nigeria) populations in the analysis. We incorporated common SNPs, gene (mRNA) expression levels, and microRNA expression levels as predictors of iGrowth. We calculated the GRM using 2.7 million HapMap imputed genotypes [28] using GCTA [129]. We calculated the GXM as a simple correlation matrix using 13,080 transcript clusters from a previous genome-wide gene expression analysis [131]. In addition we calculated a microRNA expression similarity matrix (MXM) using expression measurements for 201 microRNAs [31].

To evaluate single similarity matrix performance, we investigated the prediction performance across a set of weights. At every weight θ_{GRM} attempted ($\theta_{\text{GRM}} = 0.1, 0.2, ..., 1$ and $\theta_\epsilon = 1 - \theta_{\text{GRM}}$), the GRM did not show any predictive power (i.e. the correlation between the predicted and true iGrowth values did not differ from zero). Interestingly transcriptome data used to calculate the GXM showed an optimal prediction correlation of $R^2 = 0.38$ [0.34, 0.43] when using $\theta_{\text{GXM}} = 1$ (Figure3.3A). We estimated the 95% confidence interval of each prediction [in brackets] by 500 permutations of randomly partitioning the data into training

and test sets as described in the Methods. Furthermore, we found that for the MXM alone, the optimal prediction correlation was $R^2 = 0.35$ [0.32, 0.38] when $\theta_{MXM} = 0.4$ and $\theta_\epsilon = 0.6$ (Figure 3.3B). To further improve these results, we performed a grid search to determine whether the combination of GXM and MXM similarity matrices was able to improve iGrowth prediction. We found that prediction after performing the grid search improved resulting in a correlation of $R^2 = 0.48$ [0.45, 0.52], when $\theta_{GXM} = 0.8$, $\theta_{MXM} = 0.1$, and $\theta_\epsilon = 0.1$ (Figure 3.3C-D). The non-overlapping confidence intervals confirmed that combining genome-wide expression data improved the iGrowth predictive power of OmicKriging over using either the GXM or MXM alone.

To gauge the relative performance of OmicKriging to a baseline model, we used a method similar to the polygenic score method, by using the top gene and microRNA expression associations rather than SNP associations. We found that 255 genes and 14 microRNAs were associated with iGrowth by univariate linear regression after Bonferroni correction for multiple tests. Using these results, we performed 16-fold cross-validation to determine the top 255 genes and top 14 microRNAs using univariate linear regression in each training set. We then used the effect sizes to predict iGrowth in each test set. We also included the first 10 principal components calculated from the genotype data in each multivariate prediction model. This cross-validation procedure was repeated 500 times to generate a confidence interval. We found that the baseline model was unable to predict iGrowth, $R^2 =$ 0.0038 [-0.010, 0.064] indicating that the maximum $R^2 = 0.48$ estimated using OmicKriging represents a tremendous improvement in iGrowth prediction (Table 1).

WTCCC diseases To evaluate the prediction performance of OmicKriging for clinically relevant phenotypes, we turned to the WTCCC seven diseases cohorts. Each disease is comprised of approximately 2000 cases and 3000 shared controls. We calculated the GRM for each of the seven cohorts using approximately 400,000 directly typed SNPs. We then fit OmicKriging for each of the seven diseases. Zhou et al. had previously shown that all seven

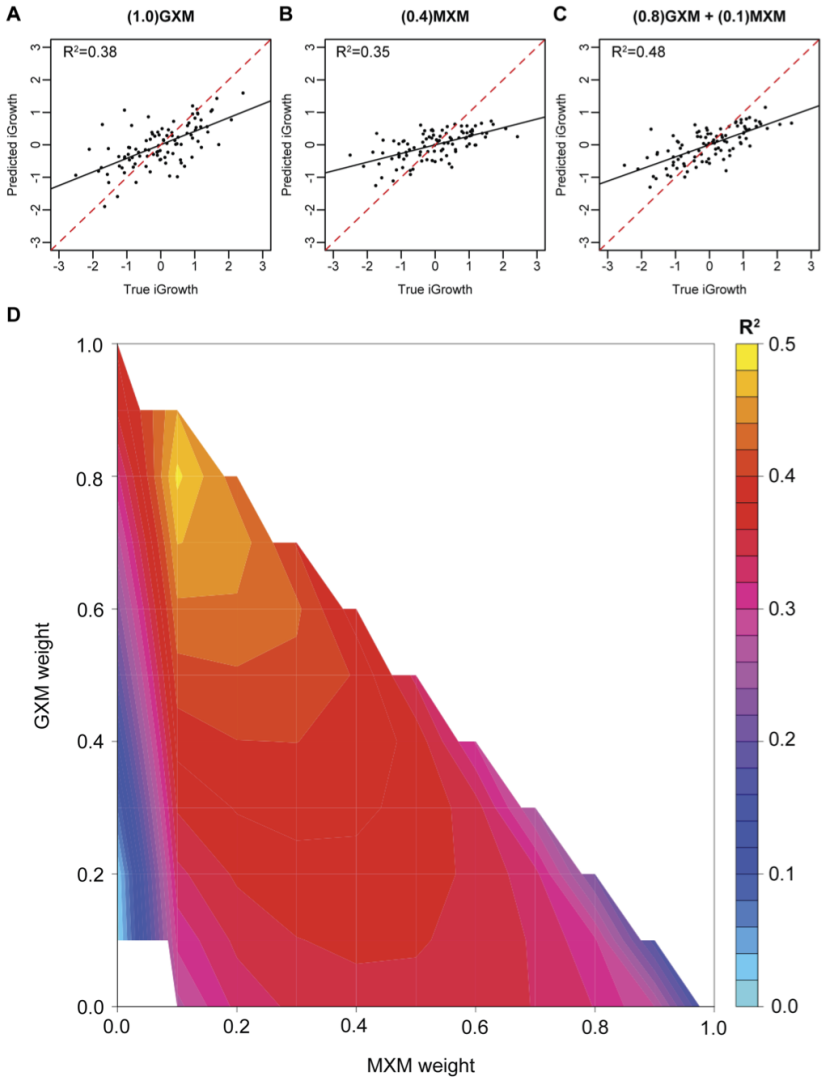

Figure 3.3: **iGrowth prediction using OmicKriging.** Predicted versus true iGrowth (n=99) using (A) the optimally weighted gene expression matrix (GXM) alone, (B) the optimally weighted microRNA expression matrix (MXM) alone, and (C) the optimally weighted combination of the two matrices from the grid search. The solid black lines represent the slopes of the regression between the predicted and true values. The red dashed lines are the identity lines representing perfect prediction (slope 1, intercept 0). (D) Results of the grid search which shows that the best iGrowth prediction correlation (R^2 = 0.48 [0.45, 0.52]) was obtained with (MXM, GXM) matrix weights of (0.1, 0.8). The R^2 values presented in the contour plot are the mean values from 500 random samplings of the data into 16 cross-validation folds.

diseases showed significant prediction performance[133]. We found in our own analysis using OmicKriging that the mean area under the ROC curve (AUC) over the seven diseases ranged from 0.598 [0.593, 0.604] when $\theta_{\mathrm{GRM}} = 0.4$ for coronary artery disease to 0.713 [0.709, 0.717] when $\theta_{\mathrm{GRM}} = 0.4$ for type 1 diabetes (Figure 3.4, Table 3).Similar to the iGrowth phenotype, the 95% confidence intervals [in brackets] were calculated through random partitions of the data 500 times into 16 subsets followed by performing cross-validated prediction on every random partition to generate a distribution of 500 AUC values. While we did perform a grid search to determine the best AUC for each WTCCC disease, optimization typically resulted in minimal improvement. That is, the optimized θ_{GRM} did not improve the AUC greater than 0.02 over the default $\theta_{\mathrm{GRM}} = 1$.

In an attempt to further improve the prediction by integrating existing information on variants from previous studies, we generated a second GRM for each disease using the SNPs within 100kb of known loci (identified outside of WTCCC studies for each disease) listed in the The National Human Genome Research Institute GWAS catalog [42] and the Database of Genotypes and Phenotypes (dbGAP) [70]. The optimal double GRM improved the predictive power of OmicKriging over using just the single common-SNP GRM alone slightly for coronary artery disease and type 2 diabetes and dramatically for Crohn's disease, rheumatoid arthritis, and type 1 diabetes (Figure 3.4). The type 1 diabetes prediction showed the largest improvement when the second GRM was added: the AUC increased from 0.713 to 0.891 [0.889, 0.892] (Figure 3.4, Table 3).

A comparison of different polygenic prediction approaches has been published by Abraham et al. [1], where the authors report that the elastic-net approach slightly outperforms other methods. We compared OmicKriging to the elastic-net penalized model and to a baseline model that uses only genome-wide significant SNPs and the first ten principal components to calculate predicted phenotypes by the polygenic score method. Both OmicKriging models (single and double GRM) outperformed the elastic-net and baseline model for

coronary artery disease and bipolar disorder (Figure 3.4, Table 3).While both OmicKriging models outperformed the baseline model for hypertension and type 2 diabetes, elastic-net performed the best for these two diseases. The OmicKriging double GRM model greatly outperforms the baseline model for Crohn's disease, rheumatoid arthritis, and type 1 diabetes. The OmicKriging double GRM model also outperforms the elastic-net model for Crohn's disease and type 1 diabetes, while elastic-net was better for rheumatoid arthritis (Figure 3.4, Table 3) .

Discussion

We propose an extension of Kriging for the Omic setting to predict complex traits which have measure high-throughput data source. We show that OmicKriging can also integrate other sources of information such as previously identified trait-associated loci or even other relevant environmental factors such as geographic proximity. We successfully translate genomic similarity into phenotypic similarity through OmicKriging. After choosing a given similarity matrix, the prediction is obtained by simply computing the weighted average of the phenotype of individuals in the training set. We show that our method is a fast, scalable and flexible approach to polygenic, and more generally poly-omic, prediction. We provide an R package called OmicKriging and a tutorial discussing how to fit OmicKriging on datasets. *http://www.scandb.org/newinterface/tools/OmicKriging.html*.

We show that by using mRNA (GXM) and microRNA (MXM) data in HapMap LCLs, we were able to predict the iGrowth phenotype with an out of sample R^2 of 0.48 with 99 samples. The use of 99 samples is highly encouraging as larger sample sizes promise potentially greater prediction performance. Remarkably, we find that the combination of mRNA-microRNA prediction R^2 was 0.10-0.13 higher than using either the GXM or MXM alone. This result is perhaps not surprising given that gene expression was shown to account for 30% of the iGrowth phenotypic variance[47]. These results suggest that biomarkers other

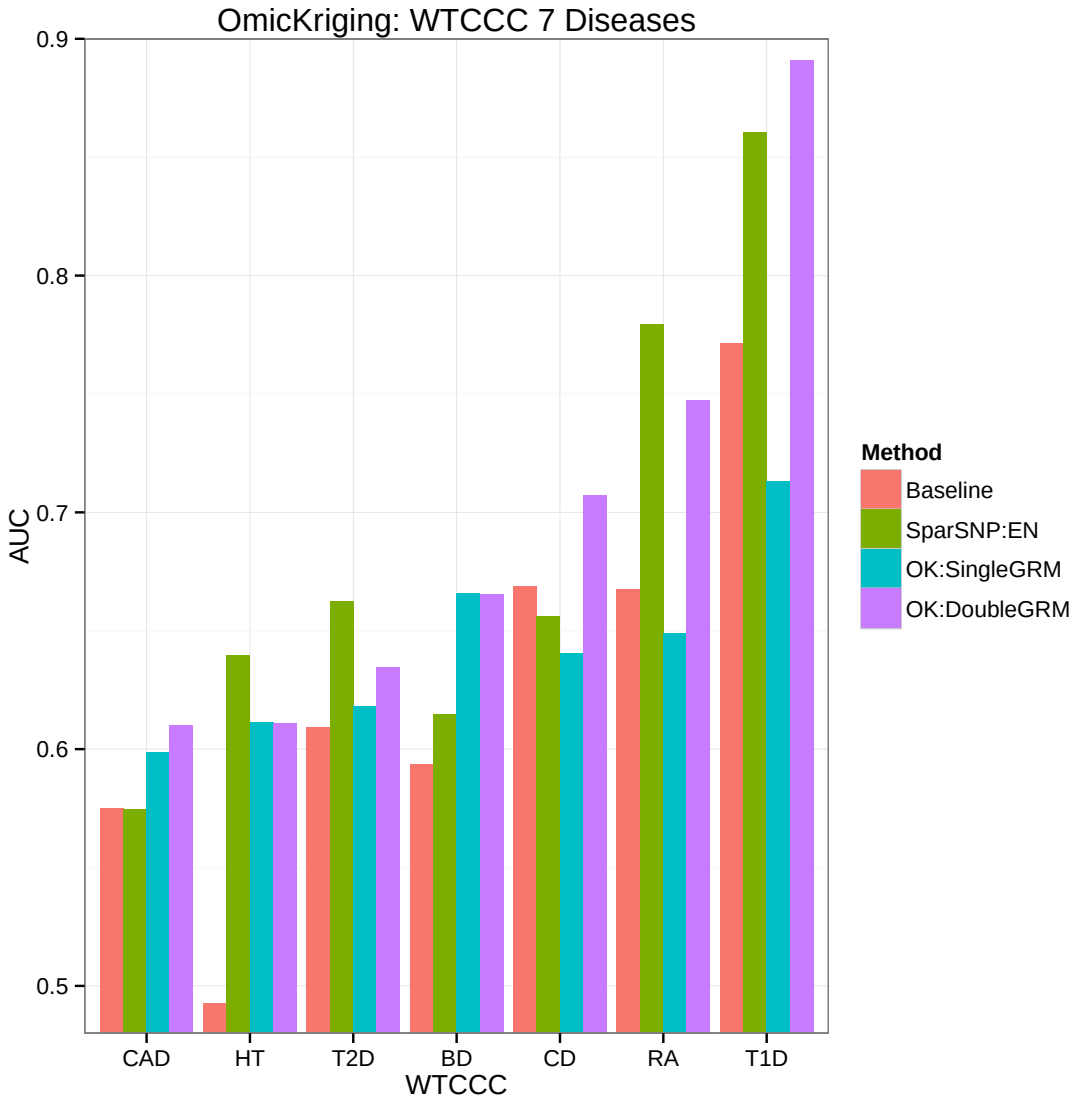

Figure 3.4: **OmicKriging prediction performance for WTCCC disease risk prediction.** Mean area under the ROC curve (AUC) for two implementations of OmicKriging for each disease from the WTCCC: a single common SNP genetic relationship matrix (OK:SingleGRM) and two optimally weighted GRMs of common SNPs and known loci (OK:DoubleGRM) for the predictions. The known loci were obtained from studies that did not include the WTCCC data to avoid over-fitting. For comparison, we also show mean AUC results of the polygenic score method using genome-wide significant loci with 10 principal components (Baseline) and the lambda-optimized elastic-net penalized model (Elastic-Net). Error bars represent the 95 % confidence intervals from multiple cross-validation runs (see Methods). BD=bipolar disorder, CAD=coronary artery disease, CD=Crohn's disease, HT=hypertension, RA=rheumatoid arthritis, T1D=type 1 diabetes, T2D=type 2 diabetes.

than SNP genotype may be useful in the clinical setting, however their utility may ultimately be highly context and environment specific.

We show successful prediction using OmicKriging for seven clinically relevant disease cohorts. Interestingly we demonstrate that OmicKriging is capable of performance similar to or better than polygenic score and similar to elastic-net under certain conditions. Importantly, OmicKriging when restricted to genotypic data performs as well as the computationally more intensive BSLMM method [133]. Therefore we feel that as data become larger in the future, our fast and scalable OmicKriging method will be more appropriate.

The average OmicKriging double GRM run time on a Xeon E5345 processor is 14 minutes, whereas the average GEMMA software [134] run time for BSLMM is 28 hours on the Xeon L5420 processor, which is a slightly newer, but comparable processor [133]. Most of the BSLMM run time is used for the Markov chain Monte Carlo (MCMC) iterations, whereas in OmicKriging, we specify the sparse effects (known GWAS loci) before the run. For the known autoimmune diseases (Crohn's disease, rheumatoid arthritis, type 1 diabetes), adding a second GRM of known loci increased the prediction AUC values over a single common SNP GRM alone to AUC values slightly higher (not significant) than those obtained by BSLMM (Table 3). [133]. These autoimmune diseases are known to have multiple associated loci of relatively strong effects, so this prior knowledge was used to improve prediction performance [114]. Unless replicated in an independent study, the results from the WTCCC data were not used to select top SNPs to avoid overfitting the data.

In the double GRM OmicKriging model we have used all SNPs to build the first similarity matrix and a subset of SNPs with prior evidence of association to build the second similarity matrix. The SNPs that were previously implicated are used in both similarity matrices. The prediction is not affected by this choice. The main reason we use this approach is to make the underlying model directly comparable to the Bayesian Sparse Linear Mixed Model, BSLMM [133], where the effect sizes of all SNPs can be represented as a mixture of two distributions:

one with small effect sizes for all SNPs and one with large effect sizes for a subset of SNPs. This way the variance explained by the second GRM is directly comparable to the PGE value (the proportion of variance explained by the sparse terms) in BSLMM. The total variance explained by the known variants will be the sum of the PGE and the proportion of the variance explained by the small effect sizes of the known variants from the first GRM.

While we recognize that assuming that a binary trait is continuous is not statistically optimal, we do so here in our initial modeling for computational reasons, as have others [60, 133]. Linear probability models for binary outcomes are considered to be adequate approximations when the proportion of cases (and controls) exceed 25% given the approximate linearity of the logit or probit functions near the origin [133]. It has been reported that the gain in statistical efficiency is hard to realize because of the added computational burden and consequent loss in numerical accuracy [133, 112]. Unlike our approach, Vazquez et al. used a probit link model for skin cancer incidence, but restricted their analysis to only 41K SNPs due to computer memory limitations [110]. In their dataset, the linear probability model would have been less appropriate since the proportion of cases was between 11-24%.

In conclusion, our results motivate the integration of OmicKriging in clinical prediction, utilizing genetic as well as other sources of patient information.

Methods

OmicKriging Approach

We propose to use an extension of the Kriging framework to integrate different omic data as well as prior information on function such as existing GWAS studies, eQTL information (genetic markers associated with gene expression levels), regulatory evidence such as provided by ENCODE studies, etc. We build the similarity matrix as a linear combination of the similarity matrices from each omic component where the coefficients or weights for each

component are chosen so that prediction performance is maximized. Given a similarity matrix, the usual Kriging formulas are used to compute the predicted values. Prediction is performed by randomly partitioning the samples into 16 subsets and using each subset as the testing set and the remaining 15 sets as the training set. This is repeated 500 times to assess the sampling variability. The correlation squared between the true and predicted values are used as performance measures. For binary/disease traits we use the area under the receiving operating curve. In this work, we assume a linear probability model for the disease status following Lee et al. [60] and Zhou et al. [133].

Omic Similarity Matrix For each omic dataset used for prediction we compute the corresponding similarity matrix. For convenience, we will denote the similarity matrix constructed from genetic data as GRM, the one constructed by mRNA expression profile data as GXM, and the ones constructed with microRNA expression data as MXM. We assume that the environmental component is independent across individuals and is represented by the identity matrix, $\mathbb{I}$. In the current implementation of the OmicKriging R package, we are computing the similarity matrix for genetic data by invoking the GCTA [129] software, whereas for other omic data we compute the similarity matrix directly in R. More specifically, the ij component of the GRM is computed as

$$\frac{1}{M}\sum_{l=1}^{M}\frac{(X_{il}^{G}-2p_l)(X_{jl}^{G}-2p_l)}{2p_l(1-p_l)}$$

and the ij component of the other omics data is computed as

$$\sum_{l=1}^{L}\frac{(X_{il}^{O}-\bar{X}_{i}^{O})(X_{jl}^{O}-\bar{X}_{j}^{O})}{\sqrt{\sum_k(X_{ik}^{O}-\bar{X}_{i}^{O})^2\sum_k(X_{jk}^{O}-\bar{X}_{j}^{O})^2}}$$

where i and j denote individuals, X_{il}^{G} is the number of reference alleles of individual i at marker l, p_l is the reference allele frequency of marker l, X_{il}^{O} is the level of omic marker l

(l is a dummy index and there is no one to one correspondence between genetic and omic markers indices), M is the number of genetic markers, L is the number of genes or omic markers, $\bar{X}_i^O = \sum_k X_{ik}^O/L$, and $\bar{X}_j^O = \sum_k X_{jk}^O/L$.

By using the correlation without prior centering and standardizing the gene expression and other continous omic traits, we are effectively giving more weight to the traits that have larger variance. The effects of different choices of the similarity matrix on the prediction accuracy will be investigated in future work.

Optimal similarity matrix under an additive model

A key component of the success of OmicKriging is understanding which proximity measures translate best into phenotypic similarity. The optimal similarity matrix depends on the underlying genetic and epigenetic architecture of the complex trait.

We will use an additive poly-omic model to motivate our choice of the similarity matrix. The phenotype for individual i, Y_i (a scalar) is represented as

$$Y_i = a + G_i + T_i + O_i + \cdots + \epsilon_i \tag{3.1}$$

where

- a is a constant,
- $G_i = \sum_{l=1}^{M} \beta_l^G X_{il}^G$ is the additive genetic component (assumed to be known), β_l^G is the effect size of the standardized genotype X_{il}^G, and M is the total number of genetic markers,
- $T_i = \sum_{l=1}^{L} \beta_l^T X_{il}^T$ (assumed to be known), β_l^T is the effect size of the standardized gene expression X_{il}^T and L is the total number of genes,
- $O_i = \sum_{l=1}^{L'} \beta_l^O X_{il}^O$ is the additive (other) omic component (assumed to be known), β_l^O

is the effect size of the standardized omic level X^O_{il}, and L' is the total number of omic markers,

- and ϵ_i is a noise term (iid, independent and identically distributed).

For notational convenience let us define X_i without a superscript to denote all three omic data such that $X_{il} = X^G_{il}$ if $l \leq M$, $X_{il} = X^T_{i,l-M}$ if $M < l \leq M + L$ and $X_{il} = X^O_{i,l-M-L}$ if $M + L < l \leq M + L + L'$ and similarly for coefficients β's such that $\beta_l = \beta^G_l$ if $l \leq M$, $\beta_l = \beta^T_{l-M}$ if $M < l \leq M + L$ and $\beta_l = \beta^O_{l-M-L}$ if $M + L < l \leq M + L + L'$.

We assume a random effects model for the β's. For convenience we also assume that the X's have been centered and standardized. If we further assume that the betas are independent, i.e. that $\text{cov}(\boldsymbol{\beta}) = \sigma^2_\beta \mathbb{I}$ then the covariance matrix of the n-vector Y will have components

$$\Sigma_{i,j} \quad = \quad \theta_G \sum_{l=1}^{M} X^G_{il} X^G_{jl} + \theta_T \sum_{l=1}^{L} X^T_{il} X^T_{jl} + \theta_O \sum_{l=1}^{L'} X^O_{il} X^O_{jl} + \theta_\epsilon \delta_{ij} \tag{3.2}$$

where δ_{ij} is the kronecker delta (1 if $i = j$ and 0 otherwise) and θ_G, θ_T, and θ_O are non negative. If all modeling assumptions were met and we assumed normality of the environmental term, this covariance matrix should be used as the similarity matrix to compute the best linear unbiased prediction (BLUP). However, these assumptions are quite strong and do not account for correlations of between marker effects, gene-gene interactions, gene-environment interactions, etc. Thus, we adopt a pragmatic approach in which we use a combination of the covariance matrices for each omic component but allow the weights θ to vary and pick the combination that provides the best predictive performance.

Independence assumption for all betas is clearly too restrictive. The effect size of genetic marker X^G_{il}, β^G_l, that influences gene expression level X^T_{ik} is likely to be correlated with β^T_k.

For unconstrained values of the cov(β_k,β_l) the covariance matrix has the form

$$\begin{aligned}\Sigma_{i,j} &= \theta_G \sum_{l=1}^{M} X_{ik}^G X_{jk}^G + \theta_T \sum_{l=1}^{L} X_{ik}^T X_{jk}^T + \theta_O \sum_{k=1}^{L'} X_{ik}^O X_{jk}^O + \theta_\epsilon \delta_{ij} \\ &+ \sum_{k \neq l} \text{cov}(\beta_k, \beta_l) X_{ik} X_{jl}\end{aligned}$$

In case prior expression quantitative trait loci (eQTL, genetic markers that have an effect on gene expression traits) information is available, it may be possible to restrict the non zero cross-correlation terms to known eQTL pairs (X_l^G,X_k^T). Additional restrictions in the values of the cov(β_l, β_k) must be imposed to be able to characterize them given existing data and care must be taken to preserve positive definiteness of $\mathbf{\Sigma}$ (all eigenvalues must be > 0). This is a complex topic that merits further research. In this paper, to keep computations within reach, we ignore the cross-correlation terms and find the coefficient thetas that maximize prediction performance.

Composite Similarity Matrix Based on the form of the optimal similarity matrix under an additive poly-omic model, we propose to use a composite similarity matrix that integrates different omic components to be used for Kriging that is a linear combination of each component similarity matrix (S_s)

$$\Sigma = \theta_1 S_1 + \theta_2 S_2 + \theta_3 S_3 + \cdots + (1 - \theta_1 - \theta_2 - \theta_3 \cdots)\mathbb{I}$$

where the weights θ's will be determined as the ones providing optimal prediction. All coefficients θ are constrained to be non-negative. The environmental component is known as the nugget term in geostatistical applications.

Kriging Formula

Within the kriging framework, the predicted phenotype of a test individual is computed as the weighted average of the phenotype of the individuals in the training set.

$$\text{Prediction}(Y_{\text{new}}) = \omega_1 Y_1 + \omega_2 Y_2 + \cdots + \omega_n Y_n \tag{3.3}$$

where the weights ω_i are a function of all $n(n+1)/2$ pairs of similarity measures. In the simplest case where no covariates are needed, the weights prescribed by the Kriging method are given by

$$\boldsymbol{\omega} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\rho} \tag{3.4}$$

where $\boldsymbol{\rho}$ is the similarity vector between the test individual and the training individuals and $\boldsymbol{\Sigma}$ is the similarity matrix of the individuals in the training set [12]. Covariates are easily included in the method by using the so called universal kriging approach [12]. Assuming there are p covariates (if only the intercept is considered, $p = 1$), let $\boldsymbol{z}$ be the p by 1 vector with covariates 1 to p corresponding to the test individual and $\mathbb{Z}$ be the n by p matrix with the p covariates for the n individuals in the training set. The weights become

$$\boldsymbol{\omega} = \Sigma^{-1}\left(\boldsymbol{\rho} + \mathbb{Z}\boldsymbol{m}\right) \tag{3.5}$$

where $\boldsymbol{m} = (\mathbb{Z}'\boldsymbol{\Sigma}^{-1}\mathbb{Z})^{-1}(\boldsymbol{z} - \mathbb{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\rho})$.

Prediction Performance

We measure prediction performance for binary traits with the area under the receiving operator characteristic curve (AUC). For quantitative traits, we use R^2, the correlation coefficient between true and predicted values squared.

Grid Search

When using a single similarity matrix, we compared the prediction performance measures for similarity matrix weights, $\theta_1 = 0, 0.1, 0.2, ...1$, and environmental component weights, $\theta_\epsilon = 1 - \theta_1$. For two similarity matrices (e.g. GRM and GXM) we allow matrix weights $\theta_1 = 0, 0.1, 0.2, ...1$ and $\theta_2 = 0, 0.1, 0.2, ...1$, with the constraint that $\theta_1 + \theta_2 \leq 1$. When two similarity matrices are used, the environmental component weights are $\theta_\epsilon = 1 - \theta_1 - \theta_2$.

Sampling Variability

To assess the uncertainty of the R^2 or AUC estimates due to sampling variability, we randomly partitioned each dataset 500 times (except elastic-net) into 16 subsets and performed cross-validation on every random partition. That is, within each cross-validation fold, 1/16 of the data was used as the test set and 15/16 of the data was used as the training set. This random sampling and cross-validation generated a distribution of 500 R^2 or AUC values for each prediction method and trait from which 95% confidence intervals were calculated (as the 0.025 and 0.975 percentiles of the distribution of R^2 or AUC values). All analyses were performed using the R statistical language and environment [103].

Genotype imputation Self-reporting Caucasian individuals from the Cholesterol and Pharmacogenetics Study [95] were genotyped on the Illumina HumanHap 300K beadchip (n=305) or the Illumina HumanHap 610K-Quad beadchip (n=282). Prior to imputation, we performed standard GWAS quality control, removing poorly called SNPs and SNPs in Hardy-Weinberg disequilibrium (HWD, $P < 0.001$). We also removed related individuals and outliers for heterozygosity or principal components. This left 562 individuals, who also had baseline and post-simvastatin treatment LDLC measurements, for imputation. Prior to imputation we pre-phased the genotype data using SHAPEIT [20] using the recommended settings. Then we used IMPUTE2 [45] to impute genotypes from the 1000 Genomes Project

[11] using the default settings for pre-phased data. A total of 8.7M SNPs with IMPUTE2-info scores > 0.3 and minor allele frequency (MAF) > 0.001 and genotypes with probabilites > 0.9 were used in the heritability estimation and prediction analyses.

Phenotype The change in low-density lipoprotein cholesterol (dLDLC) phenotype was calculated by subtracting the log-transformed mean (over the two baseline measurements) of the baseline LDLC plasma levels from the log-transformed mean (over the two visits post-treatment) of the LDLC plasma levels collected while patients were on simvastatin [95].

Expression pathway analysis To test whether the gene expression of pathways potentially related to simvastatin-induced LDLC response could predict dLDLC, we chose a few canonical pathways from the MSigDB [65] to test in our OmicKriging model as proof-of-concept for future more comprehensive pathway analyses. The pathways tested for prediction ability include BIOCARTA INFLAM PATHWAY, PID RHOA PATHWAY, PID RHOA REG PATHWAY, and REACTOME CHOLESTEROL BIOSYNTHESIS. While RHOA has been previously implicated in lipid metabolism and thus makes a plausible candidate pathway, we focus on it here over many other potential lipid metabolism pathways, because recent functional work in the CAP LCLs have revealed specific effects of RHOA (ras homolog family member A) in modulating the cholesterol-lowering effects of statin [73].

WTCCC disease analysis

We performed standard quality control for all WTCCC data sets. All WTCCC data sets were merged into a single bed file where we removed all individuals recommended by the WTCCC. This left 2937 common controls, 1868 bipolar disorder cases, 1926 cardiovascular disease cases, 1748 Crohn's disease cases, 1952 hypertension cases, 1860 rheumatoid arthritis cases, 1963 type 1 diabetes cases, and 1924 type 2 diabetes cases. We removed SNPs in HWD $P < 0.0005$ and MAF < 0.01. This resulted in approximately 388K SNPs after pruning. In

addition, we computed the GRM for all WTCCC and identified a pair of cases with unusually high relatedness that were not included in the WTCCC removal list. The duplicate individual was removed.

OmicKriging models We selected SNPs from dbGAP and NHGRI to be used in the double GRM model [42, 70]. We pruned all SNPs from studies that contained WTCCC datasets. We fit OmicKriging models with single (all SNPs) and double GRMs (all SNPs and known GWAS SNPs) in 16-fold cross-validation. We chose 16-fold because OmicKriging can be multithreaded and for these analyses, we used dual Xeon E5620 processors with 16 logical cores.

We performed a grid search (as described previously) to identify the optimal weights for single and double GRM models (Table 3). Prediction performance of all OmicKriging and baseline models was measured by area under the receiver operating characteristic (ROC) curve (AUC) with the package ROCR in R [96]. We used the R package ggplot2 [123] to generate Figure 3.4.

Polygenic score models We applied the polygenic score method by fitting the first 10 principal components and p genome-wide-significant loci jointly in 16-fold cross-validation (baseline model). Specifically, we fit Y$\sim$PC1 + $\cdots$ + PC10 + SNP1 + $\cdots$ + SNPp in each training set (15/16th of the dataset). With the remaining test set (1/16th of the dataset), the m =p+10 estimated regression coefficients ($\hat{\beta}$) are multiplied by an nxm matrix of n individuals and m principal components/genotype dosages (Z_{il}) and each individual's predicted phenotype (polygenic score) is the sum of the respective individual's products: $\sum_{l=1}^{m} Z_{il}\hat{\beta}_l$.

Elastic-net models We applied the elastic-net regularized regression method implemented by the glmnet package in R [29] to the WTCCC data. In the WTCCC data, glmnet tra-

verses a lambda penalty path for 100 iterations. We performed three replications of 16-fold cross-validation to estimate mean AUC and calculate 95% confidence intervals for prediction performance. Lambda penalty was chosen to be the value that maximized the AUC estimate for each WTCCC disease. The elastic-net mixing parameter alpha was set to 0.5.

CHAPTER 4

CLINICALLY RELEVANT PREDICTION OF BEVACIZUMAB INDUCED HYPERTENSION

Abstract

The use of genetic data to predict provocative adverse events has yet to yield clinically actionable results. This is in large part due to small cohorts ascertained during clinical trials. Nevertheless, clinical prediction of adverse events, which have primary disease analogues may benefit from the use of whole genome data available in primary diseases. Therefore we hypothesized that building a statistical learning machine that incorporates clinical trial and primary disease level data will significantly improve predictions over either dataset alone. To test this hypothesis we leveraged the CALGB 80303 and 90401 clinical trials cohorts (n=304, n=901 respectively), as well as the XC-Pleiotropy meta-analysis results for primary hypertension to build an integrative and scalable prediction model. Our results show a significant performance gain (AUC=0.72) when predicting bevacizumab-induced hypertension in CALGB 80303 when using combined predictors generated in CALGB 90401 and XCP cohorts over performance seen using each dataset separately (AUC=0.67, 0.62 respectively). This result suggests non-redundant genetic architecture between provocative and primary hypertension and motivates the use of our model in a clinical setting.

Introduction

Genome wide association (GWA) studies and more recently sequencing studies have discovered thousands of well replicated variants associated with complex phenotypes. The total phenotypic variation explained by these findings is typically well below 20% and in most cases less than 10%[72]. This low proportion of phenotypic variance explained by

trait associated loci has limited the utility of GWAS in the clinic. To assess this result, mixed effects modelling approaches were developed to estimate total variability explained by common variants captured by genotyping platforms[111, 115, 129]. These estimates have demonstrated that a substantially larger portion of the expected heritability from family studies can be explained by the collective effects of genome-wide loci over genome-wide significant loci alone[125]. This finding highlights the significant potential of whole-genome prediction over trait-associated prediction and suggests that there are clinical applications of whole-genome prediction methods which could improve patient care and realize the goals of precision medicine.

Early whole-genome prediction methods applied a polygenic score approach which uses GWAS regression coefficients estimated in training data and test genotype data to predict unobserved phenotypes by linear combination[88]. Subsequently, a variety of whole-genome prediction methods were developed after it became readily apparent that the rich and diverse field of statistical learning would have important applications in genetic medicine[17, 1]. Recent approaches have used tools that include penalized regression, Bayesian inference, kriging, bootstrap aggregation, support vector machines and deep learning. Prediction performance for traits with immune related etiology has typically been the best (AUC > 0.8), largely attributed to the HLA locus. Complex traits that are not HLA driven typically produce considerably more modest prediction performance (AUC $>$ 0.6-0.8)[120]. Nevertheless advances in whole-genome prediction tools and increasing sample sizes may produce gains significant for some complex traits[84]. Improved clinical phenotyping may also reduce noise in statistical learning applications.

While the development and application of statistical learning methodologies are critical to achieving the goal of precision medicine, significant gains may be possible through strategic selection and combination of clinically-external sources of genomic information. This may be concretely realized through applications of whole-genome prediction of drug adverse

events which have respective analogues in primary diseases. Therefore we hypothesized that whole genome prediction of drug adverse events will demonstrate a significant improvement in prediction performance through incorporation of information gained in primary disease GWAS.

To test this hypothesis we chose bevacizumab-provoked hypertension using two clinical trials cohorts, CALGB 80303[49] and 90401[53] along with the XC-Pleiotropy primary hypertension meta-analysis cohort. Bevacizumab is a humanized monoclonal antibody that inhibits VEGF induced angiogenesis and is currently used for the treatment of a variety of cancer types[67]. Hypertension is a common adverse event to bevacizumab treatment which requires careful clinical monitoring[102]. While the etiology of provocative hypertension remains unknown, it is thought to be caused by blockage of VEGF receptors as a results of decreased vasodilator nitric oxide levels (NO) levels[34]. Interestingly pre-existing hypertension does not predispose patients to grade 2/3 provocative hypertension[93]. The incidence of grade 2/3 hypertension with bevacizumab is estimated at 15%[135]. Large scale genome-wide association studies (GWAS) have examined the underlying genetic basis of hypertension but have uncovered only tens of genome-wide significant loci[24], yet twin and family studies have estimated heritability as high as 60% in men and 30% in women[56]. This finding has led researchers to characterize the genetic architecture of hypertension as highly polygenic owing little of its etiology to variants of large effect. However the genetic architecture of provocative hypertension remains poorly understood.

We built statistical learning models based on large scale whole-genome data to predict bevacizumab induced hypertension. Our approach borrows heavily from the methods developed in the statistical learning field, particularly the LASSO and Random Forests pioneered by Robert Tibshirani[104] and Leo Breiman[7] respectively. We present results which demonstrate significant prediction performance using clinical trial data and primary hypertension data separately. Finally we show that the combination of provocative and primary hyperten-

sion models trained separately results in the best overall prediction. These results suggest a non-redundant genetic architecture and should motivate the use of large scale whole-genome prediction in a clinical setting.

Results

To estimate the within-sample prediction performance of our test cohort, CALGB 90401, we fit the LASSO via 6-fold cross validation over a range of L1 pentalty parameters (λ) values for grades 2+ and 3+ hypertension. CALGB 90401 is a randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer. Our model contained 788 individuals and X SNPs. We observed that the model for grade 2+ hypertension showed relatively weak performance (AUC = 0.57)(Figure 4.1A) while our model for grade 3+ hypertension showed excellent performance (AUC 0.71)(Figure 4.2B). Interestingly prediction performance was best with penalty allowing two SNPs in the model, rs3765696 and rs742560.

To test the generalization of our grade 2+ and 3+ prediction results in 90401, we applied the LASSO coefficients estimated in CALGB 90401 to CALGB 80303, a randomized double-blind, placebo-controlled phase III trial of gemcitabine with bevacizumab versus gemcitabine with placebo. The CALGB 80303 test set consisted of patients with advanced pancreatic cancer (n=152). Generalized performance was not significant for grade 2+ hypertension(AUC = 0.5) (Figure 4.2A) however grade 3+ showed significant prediction performance (AUC = 0.6)(Figure 4.2B). This result is comparable prediction performance seen with Wellcome Trust primary hypertension using LASSO models. We found that this result was only achievable when predicting in 80303 using all coefficients estimated in the full λ path rather than just the two SNPs that performed best in 90401 alone. This suggests that the architecture of provocative hypertension may indeed possess a polygenic component.

We next aimed to improve prediction of LASSO alone by extending our approach to use

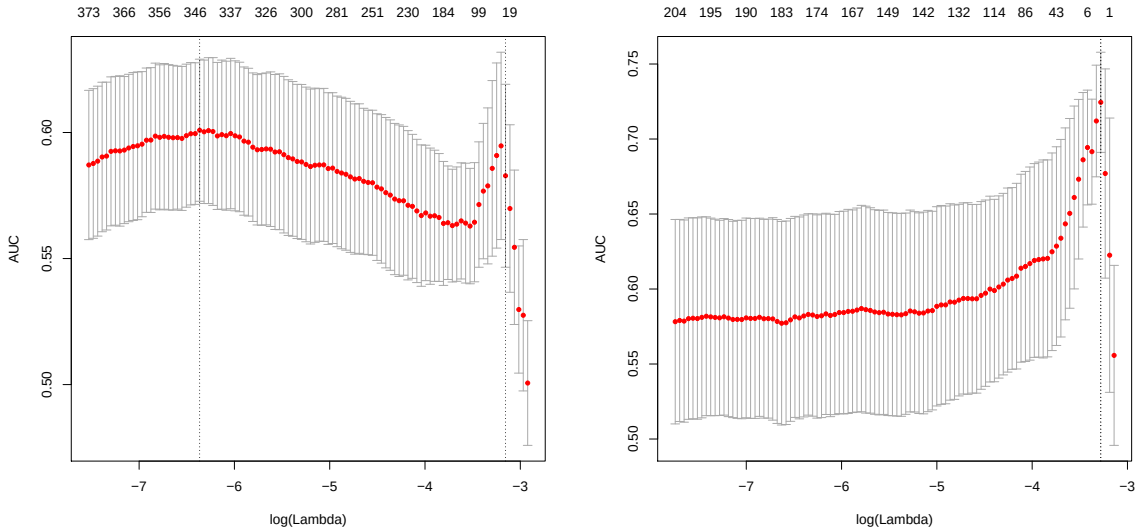

Figure 4.1: **Cross-valdiated prediction in 90401** Prediction measured by area under the ROC curve. A) CALGB 90401 prediction of hypertension grade 2+ phenotype by 6-$fold$ cross-validation. Peak prediction occurs retaining 345 SNPs in the model with an AUROC of 0.61. B) CALGB 90401 prediction of hypertension grade 3+ phenotype by 6-$fold$ cross-validation. Peak prediction occurs retaining 2 SNPs in the model with an AUROC of 0.72. package.

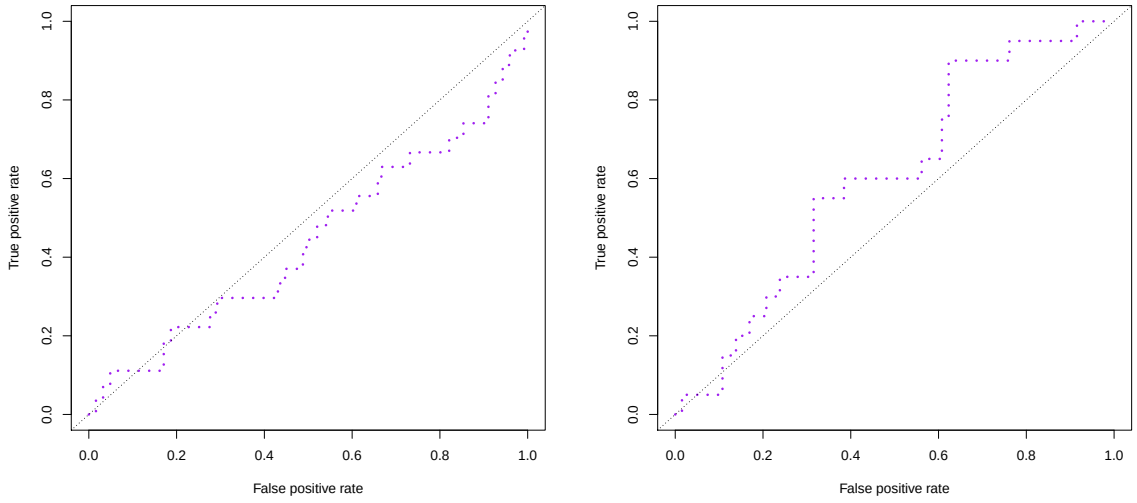

Figure 4.2: **Prediction in 80303 of grade 2+ hypertension** Grade 2+ hypertension predicted values measured by AUROC. A) 0.x AUROC of debased LASSO. B) 0.60 AUROC of debiased LASSO of. package.

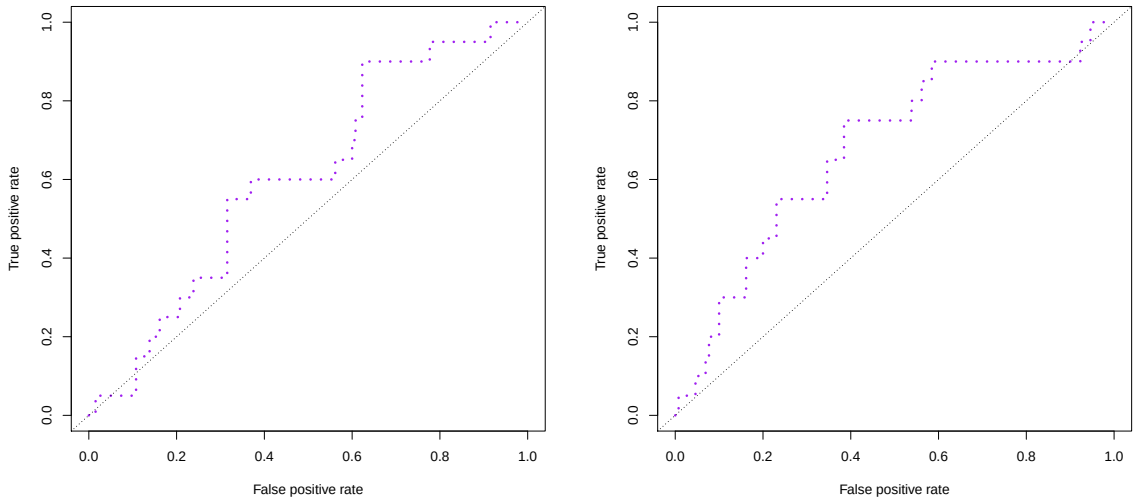

Figure 4.3: **Prediction in 80303 of grade 3+ hypertension** Grade 3+ hypertension predicted values measured by AUROC. A) 0.64 AUROC of LASSO + random forests. B) 0.68 AUROC of LASSO + random forests.

LASSO for feature selection and random forests for model training. Random forests was selected because it is a method that has both low bias and low variance estimation. Interestingly, we saw significant generalized prediction performance in 80303 (AUC = 0.64)(Figure 4.3A) which outperformed grade 2+ prediction in 90401 with the LASSO model alone. With grade 3+ hypertension, we observed significantly improved prediction performance when incorporating random forests into the LASSO model (AUC = 0.68)(Figure 4.3B).

While we were achieving significant prediction performance using provocative hypertension data alone, we hypothesized that incorporation of primary hypertension data into the provocative prediction model would significantly improve prediction performance overall. We obtained regression coefficients from a large scale meta-analysis generated by the XC-Pleiotropy consortrium and applied the polyscore approach: $\hat{y}_i = \sum_{l=1}^{m} \hat{\beta}_l X_{il}$ to CALGB 80303 data. This approach alone yields significant prediction (AUC = 0.61) comparable to prediction performance seen with Wellcome Trust primary hypertension using LASSO models.

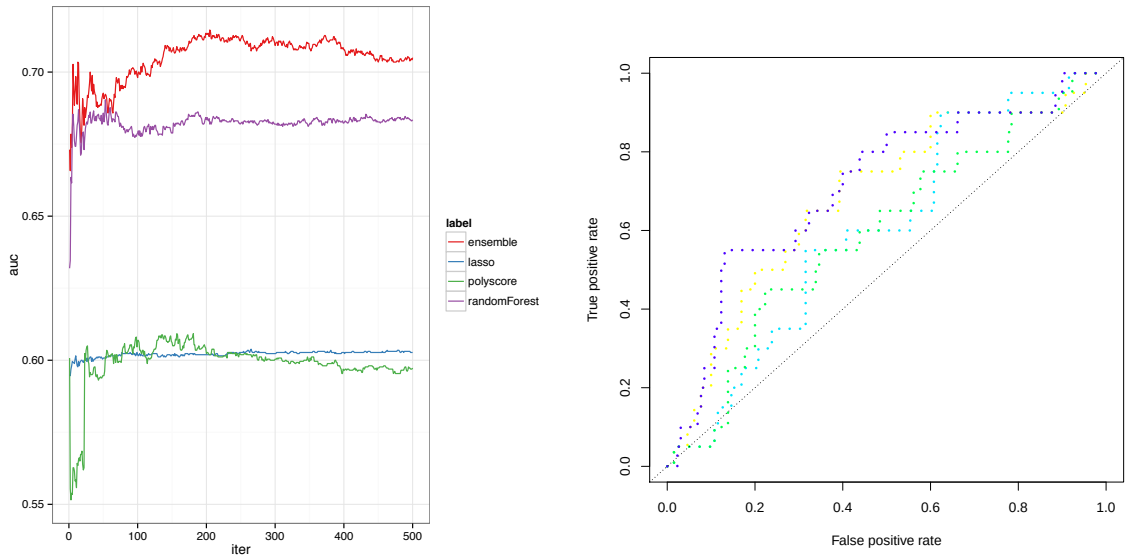

Figure 4.4: **Full Model Prediction for grade 3+** A) Prediction performance measured by area under the receiver operating characteristic curve (AUC) and colored by model tested. A legend will be included. Purple line is full model performance with an AUC of 7.1. B) Prediction performance as averaged predictions increases along the X axis.

We wanted to test whether a significant performance improvement in CALGB 80303 could be made by combining vectors of predicted response for models built using primary and provocative datasets. Therefore we produced an ensemble of predictions by taking a weighted average of both models as per the methods. We see significant performance gains (AUC=0.71)(Figure 4.4A-B) over using primary and provocative models alone. This result suggests that the genetic architecture of provocative hypertension has an underlying structure that is non-redundant with primary hypertension.

Discussion

In this study we built statistical learning models to predict bevacizumab induced hypertension. Prediction of adverse events has had little success largely owing to the small sample sizes obtained and the limitations to previous prediction methods. We were able to build several models that produced significant prediction results for bevacizumab induced hyper-

tension using two small clinical trial datasets. Our methods reveal an unanticipated degree of prediction from small clinical trial datasets. The CALGB 90401 contains only 34 cases and therefore an effective sample size of n=64. These results suggest that we may be able to identify individuals at risk for drug induced hypertension by genotype alone. At risk patients could then by better monitored for developing hypertension in the presence of therapy. Importantly our findings suggest that other provocative traits with primary complex disease analoges such as type 2 diabetes may benefit from integrative prediction approaches that we have highlighted here. In addition this method could be extended to systematically identify primary complex diseases that improve prediction performance for a given provocative phenotype.

Furthermore our approaches are scalable and computationally efficient, both of which are essential for large scale clinical implementation, particularly as clinical sample sizes continue to grow. Our computational efficiency is largely owed to the convexity of the LASSO which performs feature selection in polynomial time. While random forests is a powerful method for high variance small datasets, our prediction machine may benefit from other methods at larger sample sizes. When clinical sample sizes eventually reach the millions, deep learning methods are likely to take over as the predominant prediction approach, as significant gains for deep learning are often observed for very large samples. Nevertheless, most deep learning occurs where the sample size is much larger than the dimensionality of the features, therefore novel approaches to deep learning for underdetermined systems of equations will likely be developed. These methods will be of particular utility in the clinic when we reach millions of samples and tens of millions of features.

Materials and Methods

Quality control

A full description of the CALGB 80303 cohort and genotyping and quality control can be found as per Innocenti et al. Briefly, the Illumina HumanHap550v3 Genotyping BeadChip was used to genotype over 550,00 SNPs. The chip also contained over 7,000 SNPs located in 267 candidate genes. Imputation to HapMap 2 was done using Beagle. We removed individuals with greater than 5% SNP missingness and removed SNPs with greater than 5% missingness by individual. We pruned SNPs which exceeded expected HWE equilibrium ($p > 0.05$). This resulted in 433401 SNPs in CALGB 80303. A brief description of the CALGB 90401 cohort and genotyping can be found as per X et al. Briefly, samples were genotyped on the HumanHap610-Quad platform. We performed identical SNP quality control as per CALGB 80303 which resulted in X SNPs. Primary hypertension data used in the XC-Plieotropy is full described as per X et al. We obtained the regression coefficients estimated in 20,000 individuals and 2.6M SNPs.

Prediction Outline

Model selection and model assessment are central to inference in statistical learning. Model selection is concerned with estimating the performance of different models with goal of choosing the best one. Model assessment is concerned with how well the built models generalize to new and never before seen sets of test data. This process involves partitioning data into three discrete non-overlapping sets. The first being training data which is used to build the model and estimate model parameters. A second validation partition includes a portion of the data that are used to validate and therefore iteratively optimize the parameters of the training data. The final partition of data is the test set. This is a dataset that is never optimized on and can be thought of as a data set that is kept in a vault until the an optimized

model is reached. There are inherent limitations to an iterative approach that optimizes based on a set of validation data, namely poor generalization to the test data. Therefore over optimization in the validation set runs the risk of reduced prediction performance in the test set. While lacking a true validation data set, this study follows these principles by utilizing the XCP hypertension cohort and CALGB 90401 for training, and CALGB 80303 for testing.

Prediction Models

The LASSO approach is used to estimate an unknown fixed vector, $\beta \in \mathbb{R}^p$ given a response vector $y \in \mathbb{R}^n$ taking the form

$$y = X\beta + \epsilon \tag{4.1}$$

with $\epsilon \in \mathbb{R}^n$ as mean-zero response noise and $X \in \mathbb{R}^{n \times p}$ as the measurement matrix. Concretely, y is a vector of individual phenotypes, and X is a matrix of individual genotypes. Here we assume that the vector β is sparse such that the cardinality of the support $k = |S(\beta)|$ is $k \ll p$. The lasso is classically solved by a quadratic program of the form

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right\} \tag{4.2}$$

where λ is a regularization parameter > 0. Importantly, several authors have shown that under certain parameter conditions, the LASSO recovers the sparsity pattern asymptotically, i.e. the lasso estimator $\hat{\beta}$ is sparsistent: $P\big(S(\beta)) = S(\hat{\beta})\big) \to 1$. [116, 74, 132, 26]. We rely on this result in our application because we use the LASSO for feature selection in combination with random forests. The method random forests is an approach outlined by Leo Breiman in 2001 which uses classification and regression trees (CART) with bootstrap aggregation (bagging)[7, 21]. The method has enjoyed a great deal of success in the applied statistical

and machine learning community largely owing to low bias CARTs and low variance bagging.

We used the CALGB 90401 (n=788) as our training cohort. We fit the LASSO with k-fold cross validation with k=6 using the R package *glmnet* (Figure 1A-B).We used the regression coefficients from the max(λ_n) path to calculate the predicted response in CALGB 80303 (n=152) by

$$\hat{y}_i = \sum_{j=1}^{m} X_{ij}\hat{\beta}_{\mathrm{LASSO}_j} \tag{4.3}$$

where i is the i$^{\mathrm{th}}$ individual and j is the j$^{\mathrm{th}}$ SNP. Additionally, we fit

$$\hat{Y} \sim \mathrm{PC}_1 + \cdots + \mathrm{PC}_n \tag{4.4}$$

and used the residuals $Y - \hat{Y}$ as the vector of predicted phenotypes (Figure 2A, 3A). We measured and plotted the area under the receiver operating characteristic curve (AUROC, AUC) using the R package *ROCR*. We used the non-zero regression coefficients ($S(\hat{\beta})$) as input to the random forest model trained with the R package *randomForest*. We used default random forest parameters as we lacked a true validation dataset on which to optimize them. The R implementation of random forest assumes equal numbers of classified data, i.e. a balanced number of cases and controls. However CALGB 90401 is mostly comprised of controls with only n=34 cases. Therefore we averaged n=1000 random forest fitted models, each of the fitted models used random uniform sampling from the controls, such that each random forest fit contained n=64 individuals of 34 cases and 34 controls. Specifically the LASSO with random forest model was fit as

$$\hat{y}_i = \sum_{k=1}^{n} \mathrm{f}_{\mathrm{LASSO,RF}_k}(x) \tag{4.5}$$

where k is the k$^{\mathrm{th}}$ random forest prediction for n=1000 random forest fits. As per equation 2, we used the residuals of the predicted phenotype regressed onto 10 principal components

as the final predicted phenotype. AUROC was computed as described previously.

Our polygenic prediction approach used the regression coefficients ($\hat{\beta}$) from a meta-analysis of 20K individuals and 2.6M SNPs performed by the XC-Pleiotropy consortium to compute predicted phenotypes in CALGB 80303 (n=152). The predicted phenotypes were calculated as

$$\hat{y}_i = \sum_{l=j}^{m} X_{ij} \hat{\beta}_{XCP_j} \tag{4.6}$$

where i the i$^{\text{th}}$ individual and j is the j$^{\text{th}}$ SNP. We then regressed the predicted phenotypes on 10 principal components as per equation 2. AUROC was computed as described previously.

To integrate primary hypertension data we assumed an additive model

$$y_i = f_{\text{LASSO,RF}}(x) + g_{\text{POLYSCORE}}(x) + \epsilon \tag{4.7}$$

where $f_{\text{LASSO,RF}}(x)$ is a map from genotype to phenotype approximated using LASSO and random forest methods in CALGB 90401 (equation 3) and $g_{\text{POLYSCORE}}(x)$ is the linear combination of regression coefficients learned from XCP hypertension data (equation 4). Therefore we calculated predicted phenotype, as per equation 3 with the addition of $g_{\text{POLYSCORE}}(x)$. Specifically

$$\hat{y}_i = \sum_{k=1}^{n} \text{f}_{\text{LASSO,RF}_k}(x) + g_{\text{POLYSCORE}}(x) \tag{4.8}$$

where k is the k$^{\text{th}}$ random forest prediction for n=1000 random forest fits. described previously.

CHAPTER 5

SUMMARY

This thesis presents results of approaches to genetic medicine in the post-GWAS era. Chapter 2 explores novel distant genetic mechanisms regulating microRNA expression. Chapter 3 is focused on the extension of the Kriging method for applications in statistical genetics. Chapter 4 builds on previous statistical learning results to predict bevacizumab induced hypertension using clinical trial and primary hypertension data.

In Chapter 2, we expanded on previous eQTL studies by mapping miRQTL using an unprecedented sample size for miRQTL studies (n=5135). This resulted in the detection of 1,666 miRQTL, of which 392 were distant associations. We replicated these distant associations in the Geuvadis cohort where we saw a 26% replication rate corresponding to 67 SNPs associated with 5 microRNA. We reasoned that distant miRQTL are likely regulated through local eQTL. Therefore we tested the association of the replicated 67 loci with the expression of local transcripts. We detected 18 genes significantly associated with the 35 of the 67 loci. To further reinforce our findings, we tested the association between local mRNA transcript and distant mircroRNA level. We detected 6 transcripts significantly associated with microRNA levels. Because we hypothesized that these local transcripts were regulating distant microRNA, we performed mediation analysis using the 35 loci, 18 genes and 3 microRNA associated with the 35 loci. We found three genes were suggestive mediators of distant microRNA abundance. The only transcript that showed consistent association throughout all tests was *HEXIM1*.

HEXIM1 is a transcriptional regulator that acts to inhibit RNA polymerase II transcript elongation by sequestering the P-TEFb complex[16]. Importantly HEXIM1 is a double strand-RNA binding protein that canonically works in concert with 7SK snRNA[63]. As HEXIM1 is a key regulator of RNA polymerase II, a protein shown to elongate microRNA transcript[61], it may then be a regulator of hsa-mir-185 transcriptional machinery. Li et al.

tested whether HEXIM1 was capable of binding microRNA *in vitro*[63]. They found that the microRNA they tested (miR-16-1) could be detected in HEXIM1 immunoprecipitates. The authors did not detect the pri-mir-16-1 version in their immunoprecipitate which lead them to speculate that HEXIM1 may bind double stranded microRNA intermediates. Intermediate double stranded microRNA bound to HEXIM1 could reduce the availability of downstream mature microRNA products. Taken together, our analytical results support a negative regulatory role for HEXIM1 in modulating microRNA levels.

Furthermore, our findings support the idea that comprehensive miRQTL mapping can provide valuable insight into genetic mechanisms underlying distant regulation of miRQTL, thereby impacting on our understanding of gene regulation and associated phenotypes. This finding has the ability to expand and enrich our understanding of the mechanisms governing microRNA and mRNA transcript abundance. These results suggest further mechanistic study of their behavior and function may yield novel biological insight.

In Chapter 3 we extend and develop a method called OmicKriging, derived from the method Kriging for use in large genomic, transcriptomic and other omic data sources to interpolate and predict missing phenotypes. We show that OmicKriging is a fast and scalable framework for large scale whole-genome high-throughput prediction. We show that OmicKriging is capable of integrating external sources of information such as trait-associated-loci to adjust for the genetic architecture of the complex trait being predicted. Furthermore we show that our method is comparable and often outperforms common Bayesian approaches to whole-genome prediction. We provide an R package which makes individual application of our method straightforward.

In Chapter 4 we apply statistical learning approaches that have seen tremendous success outside the field of statistical genetics to predict bevacizumab induced hypertension in clinical trials data using separate clinical trial data and large-scale meta-analysis GWAS results of primary hypertension. We show significant prediction of hypertension using only clinical

trials data (CALGB 90401) to train a model and prediction other clinical trials data (CALGB 80303). We show significant prediction performance benefits from using the LASSO to first identify relevant factors for phenotype prediction and then learn the parameters of the model using random forests. It is however important to note that as data become very large, methods other than random forests may become more relevant for phenotype prediction. Particularly the use of deep learning which has seen large success in the fields of computer vision. However success of deep learning has primarily been with datasets of very large sample size (at least 1 million points) and modest feature space (fewer than 1 million points), therefore approaches like deep learning will require feature engineering and feature selection before they can be successfully implemented in biomedical context. This is largely owing to the fact that biomedical data tends to have a very large number of relevant features, such as the whole-genome and transcriptome, while having only few observations. Soon studies may recruit vastly larger cohort sizes, possibly in the millions, which will lessen the requirement of regularized methods such as the LASSO. We finally show that combination of data trained using CALGB 90401 and CALGB 80303 perform better than either model alone. This suggest non overlapping genetic architecture between drug-context-specific factors and primary hypertension genetic risk factors. Importantly our results should motivate the use of statistical learning approaches in the clinic to improve patient care.

APPENDIX A

FUNCTIONS

```r
'%&%' <- function(a, b) { paste(a, b, sep="") }
```

Figure A.1: **Function to concatenate strings** A simple yet essential function to concatenate strings

```r
peer.calc <- function(expr, n.fac) {
  # calculate PEER factors for m x n matrix
  # n > m
  cat("Calculating PEER factors... \n")
  require(peer)
  model = PEER()
  PEER_setPhenoMean(model,as.matrix(expr))
  dim(PEER_getPhenoMean(model))
  PEER_setNk(model,n.fac)
  PEER_getNk(model)
  PEER_update(model)

  factors = PEER_getX(model)
  rownames(factors) = rownames(expr)
  return(factors)
}
```

Figure A.2: **Function for calculating PEER factors** This function was used to calculate PEER factors in Chapter 2. This function automatically computes PEER factors for an an m $\times$ n matrix where n is much larger than m. This is ideally suited for gene expression and genotype matrices who tend to have more features than observations.

```
make.manhattan <- function(SNP.man,CHR.man,BP.man,P.man,
                           subset=NULL, colpal="Set2",
                           offset.x=0.5,
                           offset.y=0.5) {
  # make a nice manhattan plot
  require(RColorBrewer)
  require(ggplot2)
  require(plyr)

  fhs.man = data.frame(SNP=SNP.man,
                       CHR=CHR.man,
                       BP=BP.man,
                       P=-log10(P.man))
  getPalette = colorRampPalette(brewer.pal(9, "Set1"))
  color.vec = getPalette(22)
  color.dfr = data.frame(CHR=1:22, cl=color.vec)
  fhs.man = join(fhs.man, color.dfr, by = "CHR")

  manbp = fhs.man$BP
  manbp = manbp / 10^nchar(manbp)
  manbp = fhs.man$CHR + manbp
  fhs.man$BP = manbp
  fhs.man = fhs.man[order(fhs.man$BP),]

  p = ggplot(fhs.man, aes(x=BP, y=P, color=factor(BP)))
  p = p + geom_point()
  p = p +
  scale_color_manual(
                     values=as.character(fhs.man$cl))
  p = p + geom_hline(yintercept=min(fhs.man$P))
  p = p + scale_x_continuous(minor_breaks = seq(1:22),
                             breaks = seq(1:22))
  if(!is.null(subset)) {
    p = p + annotate("text",
                     x=min(fhs.man$BP)+abs((min(fhs.man$BP))*offset.x),
                     y = min(fhs.man$P)-(min(fhs.man$P)*offset.y),
                     label = subset, size =4)
  }
  p = p + ylab("-log10(p-value)")
  p = p + xlab("Chromosome")
  p = p + theme_bw()
  p = p + theme(legend.position="none")

  return(p)
}
```

Figure A.3: **Function to create Manhattan plot** This function was written for making manhattan plots in Chapter 2. The plot is based on the ggplot2 to *R* package. The function is convenient because if you supply the chromosome of the SNP and the base position, the order of the SNP-p-value results are done automatically. Futhermore there is an option to include only subsets of SNPS, for example FDR < 0.05.

```r
mediation.test <- function(mv,iv,dv,cv)
{
  # https://github.com/cran/bstats/blob/master/R/mediation.R
  # modified to include covariates
  ## mx: mediation variable
  ## iv:  indep. variable
  ## dv: dep. var.
  if(any(is.na(mv)))stop("Mediator contains missing value(s)")
  if(any(is.na(iv)))stop("Mediator contains missing value(s)")
  if(any(is.na(dv)))stop("Mediator contains missing value(s)")
  nm = length(mv); ni = length(iv); nd = length(dv);
  if(nm!=ni | nm!=nd | ni!=nd) stop("Variables have different lengths.")
  tmp = summary(lm(mv~iv));
  a = tmp$coef[2,1];sa=tmp$coef[2,2];
  tmp = summary(lm(dv~mv+iv));
  b = tmp$coef[2,1];sb=tmp$coef[2,2];
  tmp1 = b^2*sa^2+a^2*sb^2
  tmp2 = sa^2*sb^2
  zsob = a*b/sqrt(tmp1);
  psob = pnorm(-abs(zsob))*2;
  zaro = a*b/sqrt(tmp1+tmp2);
  paro = pnorm(-abs(zaro))*2;
  if(tmp1>tmp2) {
    zgm = a*b/sqrt(tmp1-tmp2)
    pgm = pnorm(-abs(zgm))*2;
  } else {
    zgm = NA
    pgm = NA;
  }
  p.value = c(psob,paro,pgm)
  z.value = c(zsob,zaro,zgm)
  out = data.frame(rbind(z.value,p.value));
  names(out) = c("Sobel","Aroian","Goodman")
  out
}
```

Figure A.4: **Function for mediation analysis** This function was adapted to include covariates in mediation analysis. The inclusion of covariates makes this function very practical.

```r
get.ncbi.snp.info <- function (results.table) {
  # annotate snps
  require(NCBI2R)
  require(plyr)
  snp.list.uniq = unique(as.character(results.table$SNP))
  rdf = GetSNPInfo(snp.list.uniq)
  colnames(rdf)[1] = "SNP"
  rdf$SNP = as.factor(rdf$SNP)
  res.anno = join(results.table, rdf, by = "SNP")
  res.anno = res.anno[order(res.anno$p.value),]
  return(res.anno)
}
```

Figure A.5: **Function for annotation of basic SNP information in R** This function was used for the annotation of basic SNP information in Chapter 2. This is an application of integrating external online databases to annotated SNP information one might be working with.

```r
gwas.anno <- function(snp.anno) {
  # annotate results with GWAS results
  require(NCBI2R)
  bb<-GetPublishedGWAS()
  out.gwas = data.frame()
  for (i in 1:length(snp.anno$SNP)) {
    cat("Search", i, "\n")
    snp.temp = snp.anno$SNP[i]
    snp.grep = bb[grep(snp.temp, bb$SNPs),]
    if (dim(snp.grep)[1] == 0) {
      tempdf = data.frame(disease="NA", gwaspval="NA")
      out.gwas = rbind(out.gwas, tempdf)
    } else {
      concat.id = paste(snp.grep$DiseaseTrait, collapse = ",")
      concat.st = paste(snp.grep$pValue, collapse = ",")
      tempdf = data.frame(disease=concat.id, gwaspval=concat.st)
      out.gwas = rbind(out.gwas, tempdf)
    }
  }
  snp.anno = cbind(snp.anno, out.gwas)
  return(snp.anno)
}
```

Figure A.6: **Function for annotation of GWAS information in R** This function was used for the annotation of GWAS SNP information in Chapter 2. This function is useful for annotating the results of a GWAS study by seamlessly connecting to external NCBI data sources.

```r
glmnet.select <- function(response, covariates, nrep.set = 11,
                          nfold.set = 10, alpha.set = 1){
  # as nrep.set goes to \infty the choice of lambda becomes very stable
  # convenience function to select best lambda over cv
  # bootstraps for model linear

  require(glmnet)
  best.lam.sim = vector()
  best.cvm.sim = vector()
  for (i in 1:nrep.set) {
    glmnet.fit = cv.glmnet(covariates, response,
                           nfolds = nfold.set, alpha = alpha.set)
    new.df = data.frame(glmnet.fit$cvm, glmnet.fit$lambda,
                        glmnet.fit$glmnet.fit$df, 1:length(glmnet.fit$lambda))
    best.lam = new.df[which.min(new.df[,1]),] # needs to be min or max
        depending
    # on cv measure (MSE min, AUC max, ...)
    cvm.best = best.lam[,1]
    nrow.max = best.lam[,4]
    best.lam.sim[i] = nrow.max
    best.cvm.sim[i] = cvm.best
  }

  cvm.avg = mean(best.cvm.sim) # average cvm
  nrow.max = as.integer(round(mean(best.lam.sim))) # best lambda over
  # cv bootstraps
  ret <- as.data.frame(glmnet.fit$glmnet.fit$beta[,nrow.max])
  ret[ret == 0.0] <- NA
  ret.vec = as.vector(ret[which(!is.na(ret)),]) # vector of non-zero betas
  names(ret.vec) = rownames(ret)[which(!is.na(ret))]
  output = list(ret.vec, cvm.avg)

  cat("avg cvm ->", cvm.avg, "lambda iteration ->", nrow.max, "with",
      length(ret.vec), "effective degrees of freedom \n")
  gc()
  return(output)
}
```

Figure A.7: **Stable selection of the regularization parameter λ in the LASSO** This function was used for the publication Gamazon et al. 2015. This function uses the *R* package *glmnet* to create a stable choice of the λ. It does this by fitting the regularized model via cross-validation n-times, and then averages the choice of λ. Here we assume that the choice of λ will remain stable as n becomes very large.

```
krigr_cross_validation <- function(corlist, pheno.df, pheno.name,
                                   Xcovamat = NULL, H2vec, nfold = 10,
                                   ncore = "all") {
  source('R/okriging.R')
  ## dependencies
  require(doMC)

  ## split into groups based on the number of cores available
  sample.ids <- pheno.df$IID
  n.samples <- length(sample.ids)

  ## detect cores
  if(ncore == "all") {
    ncore <- detectCores()
    registerDoMC(cores = ncore)
    } else {
    registerDoMC(cores = ncore)
    }

  ## set n-fold
  if(nfold == "LOOCV") {
    nfold <- n.samples
    } else if(is.numeric(nfold)) {
    nfold <- nfold
    } else if(nfold == "ncore") {
    nfold <- ncore
    } else {
    nfold <- 10
    }

  ## print core and fold numbers
  '%&%' <- function(a, b) paste(a, b, sep="")

  if(nfold == "LOOCV") {
    print('Set leave-one-out cross-validation...')
    } else {
    print('Set '%&% nfold %&%'x cross-validation...')
    }

  print('With '%&% ncore %&%' logical cores...')

  ## create groups
  groups <- 1:nfold
  rand.groups <- sample(groups, n.samples, replace=T)
  group.df <- data.frame(rand.groups, sample.ids)
  colnames(group.df) <- c("group.id", "sample.id")

  print('Running OmicKriging...')
```

Figure A.8: **Part 1 of a function which computes cross-validated OmicKriging** This is a convenience function for cross-validated multi-threaded prediction using the packing OmicKriging. This function is especially efficient as it makes uses of the specified number of cores on the user's machine.

```
## running kriging routine on each core for each testing group
time <- system.time(
res <- foreach(i = 1:nfold, .combine = rbind) %dopar% {

  ## separate test/train for round i of the cross validation
  test.set <- group.df$sample.id[group.df$group.id == i]
  train.set <- group.df$sample.id[!(group.df$sample.id %in% test.set)]

  ## run kriging
  if(!is.null(Xcovamat)) {
    okriging(idtest = test.set, idtrain = train.set, corlist = corlist,
    H2vec = H2vec, pheno = pheno.df, phenoname = pheno.name, Xcova =
        Xcovamat)
    } else {
    okriging(idtest = test.set, idtrain = train.set, corlist = corlist,
    H2vec = H2vec, pheno = pheno.df, phenoname = pheno.name)

    }
  }
)

gc()
```

Figure A.9: **Part 2 of a function which computes cross-validated OmicKriging** This is a convenience function for cross-validated multi-threaded prediction using the packing OmicKriging

```
  ## summary
  if(length(unique(res$Ytest)) == 2) {
    auc <- function(predtype, phenotype){
      require(ROCR)
      pred <- prediction(predtype, phenotype)
      perf <- performance(pred, "auc")
      aucval <- perf@y.values
      return(aucval)
      }
    print('Summary of binary phenotype...')
    print('Area under the ROC curve: '%&% auc(res$Ypred,res$Ytest) %&%'...')
    sum <- summary(glm(Ytest ~ Ypred, data = res, family = binomial))
    print(sum)
    } else {
    sum <- summary(lm(Ytest ~ Ypred, data = res))
    print(sum)
    }

  print('Finished OmicKriging in '%&% time[3] %&%' seconds')
  return(res)

}
```

Figure A.10: **Part 3 of a function which computes cross-validated OmicKriging** This is a convenience function for cross-validated multi-threaded prediction using the packing OmicKriging

```
okriging <-
    function(idtest,idtrain=NULL,corlist,H2vec,pheno,phenoname,Xcova=NULL) {
  idtest <- as.character(idtest)
  idtrain <- as.character(idtrain)
  nt <- length(idtest)
  nT <- length(idtrain)
  indall <- c(idtrain,idtest)
  if(length(unique(idtest))!=nt) warning('repeated test ids')
  if(length(unique(idtrain))!=nT) warning('repeated train ids')
  if(length(intersect(idtest,idtrain)>0)) warning('test id in training set')
  if(sum(H2vec<0) | sum(H2vec)>1) stop(' sum of weights > 1 or negative
      weights ')
  ## compute correlation matrix
  if(length(corlist)!=length(H2vec)) stop('number of correlation components
      (length(H2vec)) != number of corlist ')
  id <- diag(rep(1,nt+nT)) ## identity matrix
  Sigmall <- id * (1 - sum(H2vec))
  for(cc in 1:length(corlist)) Sigmall = Sigmall + H2vec[cc] *
      corlist[[cc]][indall,indall]
  ## row and colnames of cor should be IID
  if(sum(c(idtest,idtrain) %in% rownames(Sigmall))<(nt+nT)) stop('some
      correlations are missing')
  ## if no covariates, use intercept
  Xtest <- matrix(1,1,nt)
  Xtrain <- matrix(1,nT,1)

  if(!is.null(Xcova)) {
    Xtest <- rbind(Xtest,matrix(t(Xcova[idtest,]),ncol(Xcova),nt))
    Xtrain <- cbind(Xtrain,as.matrix(Xcova[idtrain,]))
  }

  Ytrain <- pheno[idtrain,phenoname]

  ## iSig
  iSig <- solve( Sigmall[idtrain,idtrain] )

  ctvec <- matrix(Sigmall[idtest,idtrain],nt,nT ) ## correlation between new
      id and old id (nT x nt)
  cvec <- t(ctvec)
  mtvec <- (t(Xtest) - ctvec %*% iSig %*% Xtrain) %*% solve( t(Xtrain) %*%
      iSig %*% Xtrain)

  lambt <- (ctvec + mtvec %*% t(Xtrain)) %*% iSig
  Ypred <- lambt %*% Ytrain
  Ytest <- pheno[idtest,phenoname]
  res <- data.frame(IID=idtest,Ypred,Ytest)
  rownames(res) <- idtest
  return(res)
}
```

Figure A.11: **Main OmicKriging Function** This is the primary function used for OmicKriging in Chapter 3

```r
make_grm <- function(gdsFile = NULL, grmFilePrefix = NULL, snpList = NULL,
    sampleList = NULL) {
  require(gdsfmt)
  require(SNPRelate)
  source('R/rcppcormat.r')
  source('R/grm_io.R')

  genofile <- openfn.gds(gdsFile)
  ## pull an integer dosage matrix from the GDS. Rows are samples, columns
      are SNPs, and missing values are int 3.
  X <- snpgdsGetGeno(gdsobj = genofile, sample.id = sampleList, snp.id =
      snpList, verbose = FALSE)
  ## set missing values (int 3) to properly missing
  X[X == 3] <- NA
  ## z-normalize matrix (sweep out column means, and divide out column
      standard deviations)
  X <- scale(X, center = TRUE, scale = TRUE)
  ## set missing values to new column mean, i.e. 0.0
  X[X == NA] <- 0.0
  grm <- rcppcormat(t(Xbar))

  ## pull sample IDs unless a sample list is specified
  if(is.null(sampleList)) {
    sample.ids <- sampleList
  } else {
    sample.ids <- read.gdsn(index.gdsn(genofile, "sample.id"))
  }

  ## annotate columns and rows with sample IDs
  colnames(grm) <- sample.ids
  rownames(grm) <- sample.ids

  ## write out the GRM if a file is specified
  if( !is.null(grmFilePrefix) ) {
    writeGRMBin(X = grm, prefix = grmFilePrefix)
  }

  return(grm)
}
```

Figure A.12: **Function which computes the genetic relatedness matrix** This a function which computes a genetic relatedness matrix (GRM).

```
rcppcormat <- function(snpmat){
  ## require
  require(Rcpp)
  require(RcppEigen)
  require(inline)
  ## rcpp
  crossprodCpp <- '
  using Eigen::Map;
  using Eigen::MatrixXi;
  using Eigen::Lower;
  const Map<MatrixXi> A(as<Map<MatrixXi> >(AA));
  const int           m(A.rows()), n(A.cols());
  MatrixXi          AtA(MatrixXi(n, n).setZero().
                    selfadjointView<Lower>().rankUpdate(A.adjoint()));
  MatrixXi          AAt(MatrixXi(m, m).setZero().
                              selfadjointView<Lower>().rankUpdate(A));
  return List::create(Named("crossprod(A)") = AtA,
                      Named("tcrossprod(A)") = AAt);

  '

  ## compile the cross product function
  cpcpp <- cxxfunction(signature(AA="matrix"), crossprodCpp,
  plugin="RcppEigen", verbose=FALSE)

  cormat <- cpcpp(snpmat)

  ## post work
  cormatdiag <- diag(cormat)
  cormat <- sweep(cormat, 1, cormatdiag, "/")
  cormat <- sweep(cormat, 2, cormatdiag, "/")
  return(cormat)

}
```

Figure A.13: **Function which computes a covariance matrix using RCpp** This function uses Rcpp to compute a covariance matrix which is sourced in Figure A. This function was developed because the current implementations for cross products in R are several orders slower than C++ implementations.

```r
gg_qqplot_big = function(xs, ci=0.95, subset.value=NULL, sort=TRUE,
    thin=NULL, verbose=TRUE) {
  # Fork of https://gist.github.com/slowkow/9041570.
  require(ggplot2)

  if (!is.numeric(xs)) {
    stop("data are not numeric")
  }

  N = length(xs)

  if (sort) {
    xs = sort(xs)
  }

  d.f = data.frame(observed=-log10(xs),
                   expected=-log10(1:N / N),
                   cupper=-log10(qbeta(ci,     1:N, N - 1:N + 1)),
                   clower=-log10(qbeta(1 - ci, 1:N, N - 1:N + 1)))
  gc()

  if (thin) {
    if (is.numeric(thin)) {
      d.f.top = d.f[which(d.f[,1] >= -log10(thin)),]
      d.f = d.f[which(d.f[,1] < -log10(thin)),]
      gc()

      samp.size = length(rownames(d.f.top))
      ids.thn = sample(rownames(d.f), samp.size, replace = FALSE)
      names(ids.thn) = ids.thn

      d.f = d.f[intersect(names(ids.thn), rownames(d.f)),]
      d.f = d.f[order(d.f[,1]),]
      d.f = rbind(d.f.top, d.f)
      rm(d.f.top)
      gc()
    } else {
      stop("thin is not numeric")
    }
  }
```

Figure A.14: **Big Data QQ-Plot Part 1** This is a function for creating ggplot2 quantile quantile (QQ) plots for a large number of comparisons. This may be required if for example your system cannot fit all p-values into memory to generate the plot. This function takes a vector of p-values and rather than conventionally plotting all points, it uses a threshold to plot large pvalues as a line and small p-values as points. This allows the QQ plot to be small in file size and easy to compute.

```
  log10Pe = expression(paste("Expected -log"[10], plain(P)))
  log10Po = expression(paste("Observed -log"[10], plain(P)))

  if (verbose) {
    cat("generating graphic... \n")
  }

  gc()

  p = ggplot(d.f)
  if (is.null(subset.value)) {
    p = p + geom_point(aes(expected, observed), shape=1, size=3)
  } else {
    if (!is.numeric(subset.value)) {
      stop("subset value is not numeric")
    }
    if (subset.value < min(xs) | subset.value > max(xs)) {
      stop("subset value is not within range")
    }
    llarge = -log10(subset.value)
    p = p + geom_point(data = subset(d.f,expected>llarge|observed>llarge),
                       aes(expected, observed), shape=1, size=3)
    p = p + geom_line(data = subset(d.f,expected<=llarge|observed<=llarge),
                      aes(expected, observed))
  }
  p = p + geom_abline(intercept=0, slope=1, alpha=0.5)
  p = p + geom_line(aes(expected, cupper), linetype=2)
  p = p + geom_line(aes(expected, clower), linetype=2)
  p = p + xlab(log10Pe)
  p = p + ylab(log10Po)
  p = p + theme_bw()

  return(p)
}
```

Figure A.15: **Big Data QQ-Plot Part 2** This is a function for creating ggplot2 quantile quantile (QQ) plots for a large number of comparisons. This may be required if for example your system cannot fit all p-values into memory to generate the plot. This function takes a vector of p-values and rather than conventionally plotting all points, it uses a threshold to plot large pvalues as a line and small p-values as points. This allows the QQ plot to be small in file size and easy to compute.

BIBLIOGRAPHY

[1] Gad Abraham, Adam Kowalczyk, Justin Zobel, and Michael Inouye. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic epidemiology*, 37(2):184–95, February 2013.

[2] Victor Ambros. MicroRNA pathways in flies and worms: Growth, death, fat, stress, and timing. *Cell*, 113(6):673–676, 2003.

[3] David P. Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2):281–297, 2004.

[4] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, Kenneth B. Beckman, Jianxin Shi, Rui Mei, Alexander E. Urban, Stephen B. Montgomery, Douglas F. Levinson, and Daphne Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24, 2014.

[5] Christelle Borel, Samuel Deutsch, Audrey Letourneau, Eugenia Migliavacca, Stephen B Montgomery, Antigone S Dimas, Charles E Vejnar, Homa Attar, Maryline Gagnebin, Corinne Gehrig, Emilie Falconnet, and Yann Dupre. microRNA expression levels in human fibroblasts Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. pages 68–73, 2011.

[6] a M Bowcock, J Crandall, L Daneshvar, G M Lee, B Young, V Zunzunegui, C Craik, L L Cavalli-Sforza, and M C King. Genetic analysis of cystic fibrosis: linkage of DNA and classical markers in multiplex families. *American journal of human genetics*, 39(6):699–706, 1986.

[7] Leo Breiman. Random Forrest. *Machine Learning*, pages 1–33, 2001.

[8] James C Carrington and Victor Ambros. Role of microRNAs in plant and animal development. *Science (New York, N.Y.)*, 301(5631):336–338, 2003.

[9] Tsung-Cheng Chang and Joshua T Mendell. microRNAs in vertebrate physiology and human disease. *Annual review of genomics and human genetics*, 8:215–239, 2007.

[10] Mete Civelek, Raffi Hagopian, Calvin Pan, Nam Che, Wen Pin Yang, Paul S. Kayne, Niyas K. Saleem, Henna Cederberg, Johanna Kuusisto, Peter S. Gargalovic, Todd G. Kirchgessner, Markku Laakso, and Aldons J. Lusis. Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits. *Human Molecular Genetics*, 22(15):3023–3037, 2013.

[11] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[12] Noel A C Cressie. *Statistics for Spatial Data, revised edition*, volume 928. Wiley, New York, 1993.

[13] H Dai. A Cell Proliferation Signature Is a Marker of Extremely Poor Outcome in a Subpopulation of Breast Cancer Patients. *Cancer research*, 65(10):4059–4066, May 2005.

[14] Ann K Daly. Genome-wide association studies in pharmacogenomics. *Nature reviews. Genetics*, 11(4):241–246, 2010.

[15] Christian Damasco, Antonio Lembo, Maria Patrizia Somma, Maurizio Gatti, Ferdinando Di Cunto, and Paolo Provero. A signature inferred from Drosophila mitotic genes predicts survival of breast cancer patients. *PLoS One*, 6(2):e14737, 2011.

[16] Sonja a Dames, André Schönichen, Antje Schulte, Matjaz Barboric, B Matija Peterlin, Stephan Grzesiek, and Matthias Geyer. Structure of the Cyclin T binding domain of Hexim1 and molecular basis for its recognition of P-TEFb. *Proceedings of the National Academy of Sciences of the United States of America*, 104(36):14312–14317, 2007.

[17] Gustavo de los Campos, Daniel Gianola, and David B Allison. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature reviews. Genetics*, 11(12):880–886, 2010.

[18] Gustavo de los Campos, Daniel Gianola, and Guilherme J M Rosa. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J Anim Sci*, 87(6):1883–1887, 2009.

[19] Gustavo de los Campos, Daniel Gianola, Guilherme J M Rosa, Kent A Weigel, and José Crossa. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res Camb*, 92(4):295–308, August 2010.

[20] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nat Methods*, 9(2):179–181, 2011.

[21] Misha Denil, David Matheson, and Nando De Freitas. Narrowing the Gap: Random Forests In Theory and In Practice. *Proceedings of The 31st International Conference on Machine Learning*, (1998):665–673, 2014.

[22] Sven Dennerlein, Agata Rozanska, Mateusz Wydro, Zofia M a Chrzanowska-Lightowlers, and Robert N Lightowlers. Human ERAL1 is a mitochondrial RNA chaperone involved in the assembly of the 28S small mitochondrial ribosomal subunit. *The Biochemical journal*, 430(3):551–558, 2010.

[23] Stacey L. Edwards, Jonathan Beesley, Juliet D. French, and M. Dunning. Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics*, 93(5):779–797, 2013.

[24] Georg B. Ehret. Genome-wide association studies: Contribution of genomics to understanding blood pressure and essential hypertension. *Current Hypertension Reports*, 12(1):17–25, 2010.

[25] David M Evans, Peter M Visscher, and Naomi R Wray. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human molecular genetics*, 18(18):3525–31, September 2009.

[26] Li Runze Fan Jianqing. Variable Selection via Nonconcave Penalized. 96(456):1348–1360, 2001.

[27] Ronald A Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433, 1918.

[28] Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, and Others. A second generation human haplotype map of over 3.1 million {SNP}s. *Nature*, 449(7164):851–861, 2007.

[29] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[30] Eric R. Gamazon, Dana Ziliak, Hae Kyung Im, Bonnie Lacroix, Danny S. Park, Nancy J. Cox, and R. Stephanie Huang. Genetic architecture of microRNA expression: Implications for the transcriptome and complex traits. *American Journal of Human Genetics*, 90(6):1046–1063, 2012.

[31] Eric R Gamazon, Dana Ziliak, Hae Kyung Im, Bonnie LaCroix, Danny S Park, Nancy J Cox, and R Stephanie Huang. Genetic architecture of micro{RNA} expression: Implications for the transcriptome and complex traits. *Am. J. Hum. Genet.*, 90(6):1046–1063, 2012.

[32] Daniel Gianola, Rohan L Fernando, and Alessandra Stella. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173(3):1761–76, July 2006.

[33] Arthur S Goldberger. Best linear unbiased prediction in the generalized linear regression model. *J Am Stat Assoc*, 57(298):369–375, 1962.

[34] Joey P. Granger. Vascular endothelial growth factor inhibitors and hypertension: A central role for the kidney and endothelial factors? *Hypertension*, 54(3):465–467, 2009.

[35] D Habier, R L Fernando, and J C M Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

[36] J M Hall, M K Lee, B Newman, J E Morrow, L a Anderson, B Huey, and M C King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science (New York, N.Y.)*, 250(4988):1684–1689, 1990.

[37] D A Harville. Interpolation and estimation: discussion. *Statistics: An Appraisal*, pages 281–286, 1984.

[38] Josie Hayes, Pier Paolo Peruzzi, and Sean Lawler. MicroRNAs in cancer: Biomarkers, functions and therapy. *Trends in Molecular Medicine*, 20(8):460–469, 2014.

[39] Lin He, J Michael Thomson, Michael T Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W Lowe, Gregory J Hannon, and Scott M Hammond. A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, 2005.

[40] Charles R Henderson. Estimation of genetic parameters. *Ann. Math. Stat*, 21:309–310, 1950.

[41] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[42] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.*, 106(23):9362–9367, 2009.

[43] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics*, 6(2):95–108, 2005.

[44] Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[45] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5(6):e1000529, 2009.

[46] Tianxiao Huan, Jian Rong, Chunyu Liu, Xiaoling Zhang, Kahraman Tanriverdi, Roby Joehanes, Brian H. Chen, Joanne M. Murabito, Chen Yao, Paul Courchesne, Peter J. Munson, Christopher J. ODonnell, Nancy Cox, Andrew D. Johnson, Martin G. Larson, Daniel Levy, and Jane E. Freedman. Genome-wide identification of microRNA expression quantitative trait loci. *Nature Communications*, 6:6601, 2015.

[47] Hae Kyung Im, Eric R Gamazon, Amy L Stark, R Stephanie Huang, Nancy J Cox, and M Eileen Dolan. Mixed Effects Modeling of Proliferation Rates in Cell-Based Models: Consequence for Pharmacogenomics and Cancer. *PLoS Genet.*, 8(2):e1002525, 2012.

[48] Hae Kyung Im, Micahel L Stein, and Z Zhu. Semiparametric estimation of spectral density with irregular observations. *J. Am. Stat. Assoc.*, 102(478):726–735, 2007.

[49] Federico Innocenti, Kouros Owzar, Nancy L. Cox, Patrick Evans, Michiaki Kubo, Hitoshi Zembutsu, Chen Jiang, Donna Hollis, Taisei Mushiroda, Liang Li, Paula Friedman, Liewei Wang, Dylan Glubb, Herbert Hurwitz, Kathleen M. Giacomini, Howard L. McLeod, Richard M. Goldberg, Richard L. Schilsky, Hedy L. Kindler, Yusuke Nakamura, and Mark J. Ratain. A genome-wide association study of overall survival in

pancreatic cancer patients treated with gemcitabine in CALGB 80303. *Clinical Cancer Research*, 18(2):577–584, 2012.

[50] John P a Ioannidis. Have pharmacogenomics failed, or do they just need larger-scale evidence and more replication? *Circulation: Cardiovascular Genetics*, 6(4):413–417, 2013.

[51] Luc Janss, Gustavo de los Campos, Nuala Sheehan, and Daniel Sorensen. Inferences from genomic models in stratified populations. *Genetics*, 192(2):693–704, 2012.

[52] Robert J Johnston and Oliver Hobert. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature*, 426(6968):845–849, 2003.

[53] W. K. Kelly, S. Halabi, M. Carducci, D. George, J. F. Mahoney, W. M. Stadler, M. Morris, P. Kantoff, J. P. Monk, E. Kaplan, N. J. Vogelzang, and E. J. Small. Randomized, Double-Blind, Placebo-Controlled Phase III Trial Comparing Docetaxel and Prednisone With or Without Bevacizumab in Men With Metastatic Castration-Resistant Prostate Cancer: CALGB 90401. *Journal of Clinical Oncology*, 30(13):1534–1540, 2012.

[54] Jiri Kohoutek, Dalibor Blazek, and B Matija Peterlin. Hexim1 sequesters positive transcription elongation factor b from the class II transactivator on MHC class II promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(46):17349–17354, 2006.

[55] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9(1):29, 2013.

[56] T a Kotchen, J M Kotchen, C E Grim, V George, M L Kaldunski, a W Cowley, P Hamet, and T H Chelius. Genetic determinants of hypertension: identification of candidate phenotypes. *Hypertension*, 36(1):7–13, 2000.

[57] Eric C. Lai. microRNAs: Runts of the Genome Assert Themselves. *Current Biology*, 13(23):925–936, 2003.

[58] Nan M Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature reviews. Genetics*, 7(5):385–394, 2006.

[59] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter a C 't Hoen, Jean Monlong, Manuel a Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan

Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, 2013.

[60] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[61] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060, 2004.

[62] Jeffrey T. Leek and John D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):1724–1735, 2007.

[63] Qintong Li, Jeffrey J. Cooper, Gary H. Altwerger, Michael D. Feldkamp, Madeline a. Shea, and David H. Price. HEXIM1 is a promiscuous double-stranded RNA-binding protein and interacts with RNAs in addition to 7SK in cultured cells. *Nucleic Acids Research*, 35(8):2503–2512, 2007.

[64] Shaoyu Li and Yuehua Cui. Gene-centric gene-gene interaction: A model-based kernel machine method. *Annals of Applied Statistics*, 6(3):1134–1161, 2012.

[65] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database ({MS}ig{DB}) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[66] Shuibin Lin and Richard I. Gregory. MicroRNA biogenesis pathways in cancer. *Nature Reviews Cancer*, 15(6):321–333, 2015.

[67] Maartje Los, Jeanine M L Roodhart, and Emile E Voest. Target practice: lessons from phase III trials with bevacizumab and vatalanib in the treatment of advanced colorectal cancer. *The oncologist*, 12(4):443–450, 2007.

[68] Jun Lu, Gad Getz, Eric a Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo a Ferrando, James R Downing, Tyler Jacks, H Robert Horvitz, and Todd R Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005.

[69] Michael Lynch and Kermit Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152(4):1753–1766, 1999.

[70] Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, and Others. The {NCBI} db{G}a{P} database of genotypes and phenotypes. *Nat. Genet.*, 39(10):1181–1186, 2007.

[71] R Makowsky, N M Pajewski, Y C Klimentidis, A I Vazquez, C W Duarte, D B Allison, and G de los Campos. Beyond missing heritability: prediction of complex traits. *PLoS Genet.*, 7(4):e1002051, 2011.

[72] Teri a Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia a Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven a McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[73] Marisa W Medina, Elizabeth Theusch, Devesh Naidoo, Frederick Bauzon, Kristen Stevens, Lara M Mangravite, Yu-Lin Kuang, and Ronald M Krauss. RHOA Is a Modulator of the Cholesterol-Lowering Effects of Statin. *PLoS Genetics*, 8(11):e1003058, November 2012.

[74] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[75] T H Meuwissen, B J Hayes, and M E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[76] Annemieke a Michels, Alessandro Fraldi, Qintong Li, Todd E Adamson, François Bonnet, Van Trung Nguyen, Stanley C Sedore, Jason P Price, David H Price, Luigi Lania, and Olivier Bensaude. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. *The EMBO journal*, 23(13):2608–2619, 2004.

[77] Annemieke a Michels, Van Trung Nguyen, Alessandro Fraldi, Valérie Labas, Mia Edwards, François Bonnet, Luigi Lania, and Olivier Bensaude. MAQ1 and 7SK RNA interact with CDK9/cyclin T complexes in a transcription-dependent manner. *Molecular and cellular biology*, 23(14):4859–4869, 2003.

[78] Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4), 2010.

[79] J Novembre, T Johnson, K Bryc, Z Kutalik, A R Boyko, A Auton, A Indap, K S King, S Bergmann, M R Nelson, and Others. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

[80] U Ober, J F Ayroles, E A Stone, S Richards, D Zhu, R A Gibbs, C Stricker, D Gianola, M Schlather, T F C Mackay, and Others. Using whole-genome sequence data to predict quantitative trait phenotypes in {D}rosophila melanogaster. *PLoS Genet.*, 8(5):e1002685, 2012.

[81] U Ober, M Erbe, N Long, E Porcu, M Schlather, and H Simianer. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics*, 188(3):695–708, 2011.

[82] Kathryn a O'Donnell, Erik a Wentzel, Karen I Zeller, Chi V Dang, and Joshua T Mendell. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, 435(7043):839–843, 2005.

[83] Rika Ouchida, Masatoshi Kusuhara, Noriaki Shimizu, Tetsuya Hisada, Yuichi Makino, Chikao Morimoto, Hiroshi Handa, Fumitaka Ohsuzu, and Hirotoshi Tanaka. Suppression of NF-kappaB-dependent gene expression by a hexamethylene bisacetamide-inducible protein HEXIM1 in human vascular smooth muscle cells. *Genes to cells : devoted to molecular & cellular mechanisms*, 8(2):95–107, 2003.

[84] Ju-Hyun Park, Sholom Wacholder, Mitchell H Gail, Ulrike Peters, Kevin B Jacobs, Stephen J Chanock, and Nilanjan Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*, 42(7):570–5, July 2010.

[85] Leopold Parts, Åsa K. Hedman, Sarah Keildson, Andrew J. Knights, Cei Abreu-Goodger, Martijn van de Bunt, José Afonso Guerra-Assunção, Nenad Bartonicek, Stijn van Dongen, Reedik Mägi, James Nisbet, Amy Barrett, Mattias Rantalainen, Alexandra C. Nica, Michael a. Quail, Kerrin S. Small, Daniel Glass, Anton J. Enright, John Winn, Panos Deloukas, Emmanouil T. Dermitzakis, Mark I. McCarthy, Timothy D. Spector, Richard Durbin, and Cecilia M. Lindgren. Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genetics*, 8(5), 2012.

[86] Brandon L. Pierce, Lin Tong, Lin S. Chen, Ronald Rahaman, Maria Argos, Farzana Jasmine, Shantanu Roy, Rachelle Paul-Brutus, Harm-Jan Westra, Lude Franke, Tonu Esko, Rakibuz Zaman, Tariqul Islam, Mahfuzar Rahman, John a. Baron, Muhammad G. Kibriya, and Habibul Ahsan. Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. *PLoS Genetics*, 10(12):e1004818, 2014.

[87] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel a R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575, 2007.

[88] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

[89] D R Rhodes, J Yu, K Shanker, N Deshpande, R Varambally, D Ghosh, T Barrette, A Pandey, and A M Chinnaiyan. Large-scale meta-analysis of cancer microarray data

identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9309, 2004.

[90] George K Robinson. That {BLUP} is a good thing: The estimation of random effects. *Statist Sci*, 6(1):15–32, 1991.

[91] Andreas Rosenwald, George Wright, Adrian Wiestner, Wing C Chan, Joseph M Connors, Elias Campo, Randy D Gascoyne, Thomas M Grogan, HKonrad Muller-Hermelink, Erlend B Smeland, Michael Chiorazzi, Jena M Giltnane, Elaine M Hurt, Hong Zhao, Lauren Averett, Sarah Henrickson, Liming Yang, John Powell, Wyndham H Wilson, Elaine S Jaffe, Richard Simon, Richard D Klausner, Emilio Montserrat, Francesc Bosch, Timothy C Greiner, Dennis D Weisenburger, Warren G Sanger, Bhavana J Dave, James C Lynch, Julie Vose, James O Armitage, Richard I Fisher, Thomas P Miller, Michael Leblanc, German Ott, Stein Kvaloy, Harald Holte, Jan Delabie, and Louis M Staudt. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3(2):185–197, February 2003.

[92] D T Ross, U Scherf, M B Eisen, C M Perou, C Rees, P Spellman, V Iyer, S S Jeffrey, M de Rijn, and M Waltham. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3):227–235, 2000.

[93] M Wasif Saif. Managing bevacizumab-related toxicities in patients with colorectal cancer. *The journal of supportive oncology*, 7(6):245–251, 2009.

[94] Noriaki Shimizu, Rika Ouchida, Noritada Yoshikawa, Tetsuya Hisada, Hajime Watanabe, Kensaku Okamoto, Masatoshi Kusuhara, Hiroshi Handa, Chikao Morimoto, and Hirotoshi Tanaka. HEXIM1 forms a transcriptionally abortive complex with glucocorticoid receptor without involving 7SK RNA and positive transcription elongation factor b. *Proceedings of the National Academy of Sciences of the United States of America*, 102(24):8555–8560, 2005.

[95] J A Simon, F Lin, S B Hulley, P J Blanche, D Waters, S Shiboski, J I Rotter, D A Nickerson, H Yang, M Saad, and Others. Phenotypic predictors of response to simvastatin therapy among {A}frican-{A}mericans and {C}aucasians: the Cholesterol and Pharmacogenetics ({CAP}) Study. *Am J Cardiol*, 97(6):843, 2006.

[96] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. {ROCR}: {v}isualizing classifier performance in {R}. *Bioinformatics*, 21(20):3940–3941, 2005.

[97] Harris S Soifer, John J Rossi, and Pal Saetrom. MicroRNAs in Disease and Potential Therapeutic Applications. *Mol Ther*, 15(12):2070–2079, September 2007.

[98] M H W Starmans, B Krishnapuram, H Steck, H Horlings, D S A Nuyten, M J van de Vijver, R Seigneuric, F M Buffa, A L Harris, B G Wouters, and P Lambin. Ro-

bust prognostic value of a knowledge-based proliferation signature across large patient microarray studies spanning different cancer types. *British Journal of Cancer*, 99(11):1884, November 2008.

[99] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, 6(5):1–11, 2010.

[100] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–7, 2012.

[101] Michael Leonard Stein. *Interpolation of spatial data: some theory for kriging*. Springer Verlag, 1999.

[102] Kostas N. Syrigos, Eleni Karapanagiotou, Paraskevi Boura, Christian Manegold, and Kevin Harrington. Bevacizumab-induced hypertension: Pathogenesis and management. *BioDrugs*, 25(3):159–169, 2011.

[103] R Development Team. R: A language and environment for statistical computing. Technical report, ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project.org, 2005.

[104] Robert Tibshirani. Regression Selection and Shrinkage via the Lasso, 1994.

[105] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, January 1996.

[106] Chao Tu, Xiaomei Zhou, Sergey G Tarasov, Joseph E Tropea, Brian P Austin, David S Waugh, Donald L Court, and Xinhua Ji. The Era GTPase recognizes the GAUCAC-CUCC sequence and binds helix 45 near the 3' end of 16S rRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10156–10161, 2011.

[107] Chao Tu, Xiaomei Zhou, Joseph E Tropea, Brian P Austin, David S Waugh, Donald L Court, and Xinhua Ji. Structure of ERA in complex with the 3' end of 16S rRNA: implications for ribosome biogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(35):14843–14848, 2009.

[108] Marieke van Kouwenhove, Martijn Kedde, and Reuven Agami. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nature reviews. Cancer*, 11(9):644–656, 2011.

[109] P M VanRaden. Efficient methods to compute genomic predictions. *J Dairy Sci*, 91(11):4414–4423, 2008.

[110] Ana I Vazquez, Gustavo de los Campos, Yann C Klimentidis, Guilherme J M Rosa, Daniel Gianola, Nengjun Yi, and David B Allison. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. *Genetics*, 192(4):1493–502, December 2012.

[111] Anna a E Vinkhuyzen, Naomi R Wray, Jian Yang, Michael E Goddard, and Peter M Visscher. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet*, 47:75–95, 2013.

[112] P M Visscher, C S Haley, and S A Knott. Mapping QTLs for binary traits in backcross and F2 populations. *Genetics research*, 68(01):55, August 1996.

[113] Peter M. Visscher, Matthew a. Brown, Mark I. McCarthy, and Jian Yang. Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1):7–24, 2012.

[114] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of {GWAS} discovery. *Am. J. Hum. Genet.*, 90(1):7–24, 2012.

[115] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era–concepts and misconceptions. *Nature reviews. Genetics*, 9(4):255–266, 2008.

[116] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ???1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[117] Lucas D. Ward and Manolis Kellis. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1):930–934, 2012.

[118] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):1001–1006, 2014.

[119] Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, Alexandra Zhernakova, Daria V. Zhernakova, Jan H. Veldink, Leonard H. Van den Berg, Juha Karjalainen, Sebo Withoff, André G. Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter a. C. 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Michael a. Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina a. Gharib, Daniel a. Enquobahrie, Thomas Lumley, Grant W. Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C. Jansen, Peter M. Visscher, Julian C. Knight, Bruce M. Psaty, Samuli Ripatti, Alexander Teumer, Timothy M.

Frayling, Andres Metspalu, Joyce B. J. van Meurs, and Lude Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013.

[120] Heather E Wheeler, Keston Aquino-michaels, Eric R Gamazon, Vassily V Trubetskoy, M Eileen Dolan, R Stephanie Huang, Nancy J Cox, and Hae Kyung Im. Poly-Omic Prediction of Complex Traits : OmicKriging. 38(5):402–415, 2014.

[121] M Whirl-Carrillo, E M McDonagh, J M Hebert, L Gong, K Sangkuhl, C F Thorn, R B Altman, and T E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics*, 92(4):414–7, 2012.

[122] M L Whitfield, L K George, G D Grant, and C M Perou. Common markers of proliferation. *Nature Reviews Cancer*, 6(2):99–106, 2006.

[123] Hadley Wickham. *ggplot2: {e}legant graphics for data analysis*. Springer Publishing Company, Incorporated, 2009.

[124] Bryan M Wittmann, Koh Fujinaga, Huayun Deng, Ndiya Ogba, and Monica M Montano. The breast cell growth inhibitor, estrogen down regulated gene 1, modulates a novel functional interaction between estrogen receptor alpha and transcriptional elongation factor cyclin T1. *Oncogene*, 24(36):5576–5588, 2005.

[125] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, ..., and Timothy M. Frayling. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, In Press(11), 2014.

[126] Fred a Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, Abdel Abdellaoui, Sandra Batista, Casey Butler, Guanhua Chen, Ting-Huei Chen, David D'Ambrosio, Paul Gallins, Min Jin Ha, Jouke Jan Hottenga, Shunping Huang, Mathijs Kattenberg, Jaspreet Kochar, Christel M Middeldorp, Ani Qu, Andrey Shabalin, Jay Tischfield, Laura Todd, Jung-Ying Tzeng, Gerard van Grootheest, Jacqueline M Vink, Qi Wang, Wei Wang, Weibo Wang, Gonneke Willemsen, Johannes H Smit, Eco J de Geus, Zhaoyu Yin, Brenda W J H Penninx, and Dorret I Boomsma. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–7, 2014.

[127] Sewall Wright. Systems of mating. Parts {I-V}. *Genetics*, 6(2):111–178, 1921.

[128] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

[129] Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82, 2011.

[130] Jasper H N Yik, Ruichuan Chen, Andrea C. Pezda, and Qiang Zhou. Compensatory contributions of HEXIM1 and HEXIM2 in maintaining the balance of active and inactive positive transcription elongation factor b complexes for control of transcription. *Journal of Biological Chemistry*, 280(16):16368–16376, 2005.

[131] Wei Zhang, Shiwei Duan, Emily O Kistner, Wasim K Bleibel, R Stephanie Huang, Tyson A Clark, Tina X Chen, Anthony C Schweitzer, John E Blume, Nancy J Cox, and Others. Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.*, 82(3):631–640, 2008.

[132] P Zhao and B Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[133] X Zhou, P Carbonetto, and M Stephens. Polygenic Modeling with {B}ayesian Sparse Linear Mixed Models. *PLoS Genet.*, 9(2):e1003264, 2013.

[134] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, 44(7):821–824, 2012.

[135] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–9, 2014.

[136] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, January 2005.